SOMATIC VARIATION AND GENOMIC INSTABILITY GENERATED BY RETROTRANSPOSONS

by

Tara Theresa Doucet-O'Hare

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore. MD

December, 2015

ABSTRACT

SOMATIC VARIATION AND GENOMIC INSTABILITY GENERATED BY RETROTRANSPOSONS

Tara Theresa Doucet-O'Hare

Haig H. Kazazian Jr. M.D., Advisor

Genomic instability is strongly linked to the development and malignancy of cancer and by studying premalignant conditions we can gain a better understanding of the sources of genomic instability and improve cancer prevention and treatment. The genome is very unstable in both premalignant and malignant conditions; however, it is unknown as to what extent different types of instability contribute. Retrotransposition is an active source of genomic instability in the human genome and has the potential to change DNA structure and RNA expression. Retrotransposons are repetitive sequences that can "copy and paste" into the genome at new sites within an individual cell, and hundreds are known to be active in the human genome. Despite the enormous influence of retrotransposons on the genome composition of many organisms, the degree to which they contribute to somatic genomic instability is unknown. Because retrotransposition has been observed in many gastrointestinal epithelial cancer types, we focused on L1 mobilization as a source of instability in cancer. We hypothesized that L1 retrotransposition is active in esophageal squamous cell carcinoma (SCC), esophageal adenocarcinoma (EAC), and EAC's precursor Barrett's esophagus (BE). To test our hypothesis, we evaluated 5 patients with benign BE, 5 patients with BE and concomitant esophageal adenocarcinoma (EAC), 10 additional patients with EAC, and 9 patients with SCC to determine the level of L1 activity in these diseases. Following L1-seq, we confirmed 160 somatic insertions by PCR in 17 of 29 individuals. We observed clonal amplification of several insertions which

appeared to originate in normal esophagus (NE) or BE and were later clonally expanded in BE, in EAC, or in SCC. Additionally, we observed evidence of clonality within the EAC cases: specifically, 22 of 25 EAC-only insertions were present identically in distinct regions available from the same tumor, suggesting they occurred in the founding tumor cell of these lesions. Our data show that somatic retrotransposition occurs early in many patients with BE and EAC, and indicate that early events occurring in histologically normal esophageal cells may be clonally expanded in esophageal adenocarcinogenesis. Additionally, we evaluated L1 ORF1 protein expression in 9 of the carcinoma cases for which formalin-fixed paraffin embedded tissue was available. Using immunohistochemistry, we detected expression of ORF1p in all tumors evaluated. Interestingly, we also observed dim ORF1p expression in the normal esophagus of all 4 patients for whom additional blocks of normal esophagus containing squamous epithelium was available. To determine if ORF1p expression is a hallmark of unaffected tissues, we obtained both skin and esophageal biopsies from two unaffected individuals. In both biopsies, ORF1p expression was evident in the squamous cell epithelium. ORF1p may be expressed in many normal epithelial tissues which could account for the high incidence of somatic retrotransposition events in epithelial cancers. Thus, our data show that L1 is weakly expressed in normal esophagus and retrotransposition can occur in normal tissue during the development of esophageal adenocarcinoma and squamous cell carcinoma. Due to the pervasive activity of retrotransposons in epithelial cancer, it is likely that somatic insertions may play a role in some tumorigenesis.

ACKNOWLEDGEMENTS

I must first thank my mother and father, Andrea and Kermit Doucet for instilling within me a good work ethic and a thirst for knowledge. Without such intelligent, kind, and exemplary parents, I highly doubt I would be the person I am today. I must also acknowledge my college mentors from Clemson University, Dr. Kerry Smith and Dr. Cheryl Smith, without whom I would never have considered a PhD in genetics nor believed I could succeed as a scientist. Furthermore, I must thank my mentor, Haig Kazazian, who is a unique mentor and a personality I will never forget. With Haig's guidance, I have become an expert in the field of mobile elements and have had the opportunity to collaborate with many colleagues whose advice and criticisms have been invaluable. I know I was truly fortunate to benefit from the wisdom of such a revered expert in our field. Additionally, I must acknowledge the love and support of my friends, classmates, and colleagues, too numerous to list here, without whom I would not be the scientist I am today. Finally, I would like to dedicate my thesis to my husband, Steven O'Hare who has staunchly supported me through my journey in graduate school.

TABLE OF CONTENTS

List of Tables

List of Figures:

CHAPTER 1:

Introduction

Transposable elements (TE): diversity, mechanism, and contribution to human evolution

The human genome is a mystery which is fast being unraveled with the advent of new technologies and techniques for ascertaining its secrets. One of the most surprising and interesting findings concerning the genome is the proportion of it which is made up of repetitive elements. In 1968, Britten and Kohne were the first to observe the highly repetitive nature of the genome using DNA hybridization kinetics (1). It is now known that the repeats comprising the genome expanded its size due to the mobility of elements first discovered in maize (2). McClintock observed two elements she named Activator (*Ac*) and Dissociator (*Ds*) that were able to transpose (i.e. transposable elements) from one chromosomal location to another and thereby affect the phenotype of the organism at sites termed mutable loci (2,3). Two decades transpired before McClintock's work was widely accepted; however, following its acceptance there have been thousands of publications throughout the scientific community concerning transposable elements in genomic DNA. In fact, the seminal papers describing the sequencing of the human genome revealed that roughly 45 percent of its sequence was composed of interspersed repeats(4,5). Transposable elements are found in all domains of life including Bacteria (6), Archaea (7), and Eukarya (2) and are characterized and classified based upon the mechanism by which they propagate.

DNA transposons are the first class of transposable elements and mobilize via a "cut-and-paste" mechanism during which the element excises itself from its current location and re-inserts into a new one. These elements have a propensity to insert into a new location which is proximal to their progenitor locus (8) and many of the elements are site specific (9). There are many families of DNA transposons scattered among various species with variations in the chemical process of the transposition reaction; however, the critical steps are common to all elements.

Initially the 3' hydroxyl groups are exposed at the transposon ends of the donor site, then a strand-transfer reaction occurs to integrate the element into the target site (9,10). The strand transfer occurs via a nucleophilic attack on the target site by an exposed 3' hydroxyl group. In addition to this more common mechanism, there are two other mechanisms employed by different DNA transposons. Helitrons (11), a DNA transposon found exclusively in eukaryotes that replicates via a rolling circle mechanism, and Politons/ Mavericks (12,13) encode proteins responsible for their self-propagation via an integrase-dependent pathway.

The next two classes of transposable elements are the so-called "retrotransposons" named for their ability to utilize a reverse transcriptase and mobilize through an RNA intermediate. The retrotransposons flanked by long terminal repeats (LTR retrotransposons) are found in many organisms from yeast (14), to insects (15). and to mice (16). LTR retrotransposons have an internal promoter in the 5' repeat and their proteins are translated in the cytoplasm of the cell where their proteins and RNA form a virus-like-particle (VLP). The VLP then enters the nucleus and performs reverse transcription. The internal structure of LTRs is diverse which may reflect an evolutionary history of multiple time-points when the infectious retroviruses were unable to leave an infected cell, potentially due to their inability to form infectious particles. LTRs comprise approximately 8% of the human genome (4) and are thought to be largely transcriptionally inactive with the potential exception of HERV-K (17).

The final group of retrotransposons is comprised of non-LTR retrotransposons, the only transposable elements known to be active in the human genome (see Figure 1.1). These elements lack long terminal repeats and seem to have an evolutionary connection to group II introns (18,19) as they have similar mechanisms in their life cycle. The elements also have shared history with telomeres because in Drosophila, telomeres are comprised of long tandem

arrays of two non LTR- retrotransposons (20). These two types of retrotransposons, *Het-A* and *TART*, are components of a robust telomere maintenance system and were the first transposable elements shown to have a "*bona fide* role in cell structure," (20). In the human genome, the long interspersed element- 1 (LINE-1 or L1) is the only autonomous member of the family of non-LTR retrotransposons. Not only has L1 maintained its activity in *Homo sapiens*, but it also mobilizes many of the other examples of non-LTR retrotransposons in humans. Active L1 elements are approximately 6,000 nucleotides long and possess two open reading frames, ORF1 and ORF2, which contain an RNA chaperone protein and a protein with endonuclease and reverse transcriptase abilities respectively (21). L1 also has an approximately 900 nucleotide 5' untranslated region and ORF2 is followed by an approximately 200 nucleotide UTR (21) and a poly(A) tail of varying length depending on the age of the element. L1s are transcribed as poly-adenylated transcripts from an internal promoter (22,23). The ORF1 protein (ORF1p) seems to bind the L1 RNA as a trimer, and acts as a nucleic acid chaperone (24), and appears to be necessary for robust retrotransposition to occur (25–28). The ORF1 protein is 40 kD, has an N-terminal domain, coiled coil domain (CCD), an RNA recognition motif (RRM), and a C terminal domain (CTD) (29). Additionally, ORF1p demonstrates bias toward binding its own RNA (26,30,31). In recent years the structure of the ORF1p trimer has been further explored through crystallization (32); however, ORF2p has yet to be crystallized and therefore less is known about its functional mechanisms. It is known that ORF2p is 150 kD and 1278 amino acids (as coded by the L1.2A allele) and has two key functions with respect to retrotransposition (25,33–36). ORF2p is responsible for nicking a single strand of the double stranded DNA and reverse transcribing the RNA into the genome (34,36). The ORF2 protein is thought to have four main components: the endonuclease domain in the N terminal region (36), a set of seven subdomains

of typical reverse transcriptases (37), and the third and fourth domains are of unknown function including the 'Z' segment and a 3' cysteine-rich motif (25,38). Clements and Singer deleted various regions of ORF2p in an attempt to identify domains essential to protein function outside of the reverse transcriptase domain (39). The authors concluded that the octapeptide sequence and its adjacent amino acids in the Z region are essential for reverse transcriptase activity; however, the endonuclease and the cysteine rich domains are not necessary. Furthermore, translation of ORF2p is unusual because of the bicistronic nature of the L1 transcript. Translation of ORF2p appears to predicate on the presence of an upstream ORF, an inter-ORF region, and may involve an IRES in mouse (40–42).

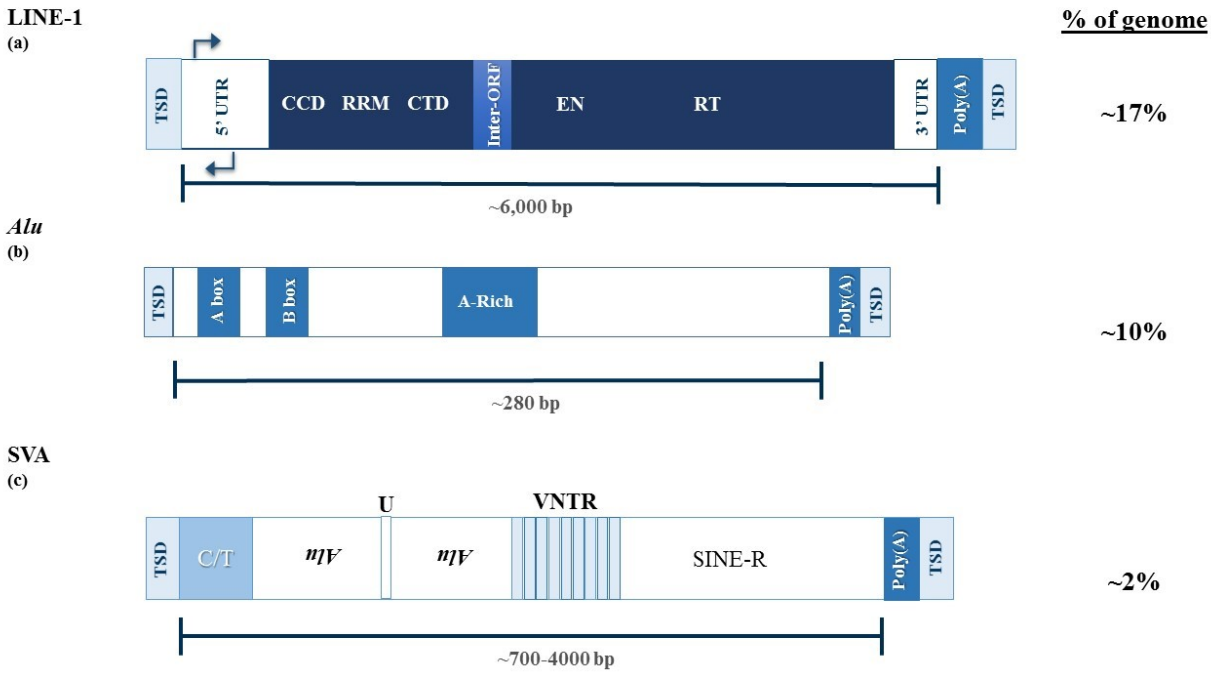


Figure 1.1: Structural characteristics of non-LTR retrotransposons. **(a)** intact LINE-1 elements are approximately 6 kb in length, and have two open reading frames (ORF1 and ORF2) which

are discussed in the text. LINE-1 elements have poly(a)-tails and are flanked by target site duplications (TSDs). The 5' UTR of the element has two promotors (indicated by the arrows) in both the sense and antisense directions. **(b)** *Alu* elements are ~280 bp in length and are rarely 5' truncated. The elements are comprised of two 7SL- derived monomers (e.g. the left and right monomers, white boxes flanking the "U"). Transcription is pol III dependent and mediated by A and B boxes indicated in the figure. The region in the center of the element is A-rich, *Alu* elements have poly(A) tails, and are flanked by TSDs. **(c)** SVA elements are highly variable in length because they include a variable nucleotide tandem repeat (VNTR). The 5' end of the element contains a CCCTCT repeat and two inverted *Alu* sequences, and the 3' end has homology to the HERV-K10 right LTR (SINE-R). SVA elements also possess poly(A) tails and have target site duplications (TSDs).

The life cycle of the L1 begins with transcription of L1 RNA in the nucleus followed by mRNA export to the cytoplasm. Next, the L1 mRNA bicistronic message is translated into the proteins ORF1 and ORF2. The L1 RNA is bound by a trimer of ORF1 (24) and together ORF1 and ORF2 form a ribonucleoprotein particle which travels back into the nucleus either during replication while the nuclear envelope is degraded, or through another currently undetermined mechanism along with other proteins and RNAs.. Once the RNP is in the nucleus, the endonuclease of ORF2 cleaves open the DNA on a single strand and the process known as target-primed reverse transcription (TPRT) begins (43). The resolution in the DNA break created by the endonuclease of ORF2 is purported to be amended by DNA repair mechanisms in the cell but this process is still largely a mystery (44). The process of TPRT also leaves target site duplications (TSDs) which reflect the nucleotide sequence between the initial nick made by the endonuclease of ORF2 and the obligate secondary break on the opposite strand of the DNA.

During the process of TPRT the single strand overhangs, which formed due to the single strand breaks on alternating strands, function as the primers in the reverse transcription reaction. Normally, one of the two overhangs anneals to the poly(A) tail of the RNA, the sequence following it is reverse transcribed using the L1 RNA as a template, and then the second strand of the DNA is resolved by DNA repair machinery which copies the newly integrated insertion on the second strand. These TSDs are considered a hallmark of the canonical method of retrotransposition (45). Occasionally during the process of TPRT, an inversion of the 5' end of the element occurs and the mechanism for this has been referred to as 'twin-priming' (45). In twin priming, one of the single stranded overhangs anneals to the poly(A) tail and the other anneals at an internal region of the same L1 mRNA creating an inversion near the 5' end of the element (45). The product of this process is an insertion where the first several hundred base-pairs of the inserted, 5' truncated L1 element sequence are in the opposite orientation to the remainder of the inserted sequence; e.g. base-pairs 5,000-4,000 are in the 3'-5' orientation while the remainder of the inserted element base-pairs 5,000-6,000 are in the 5'-3' orientation.

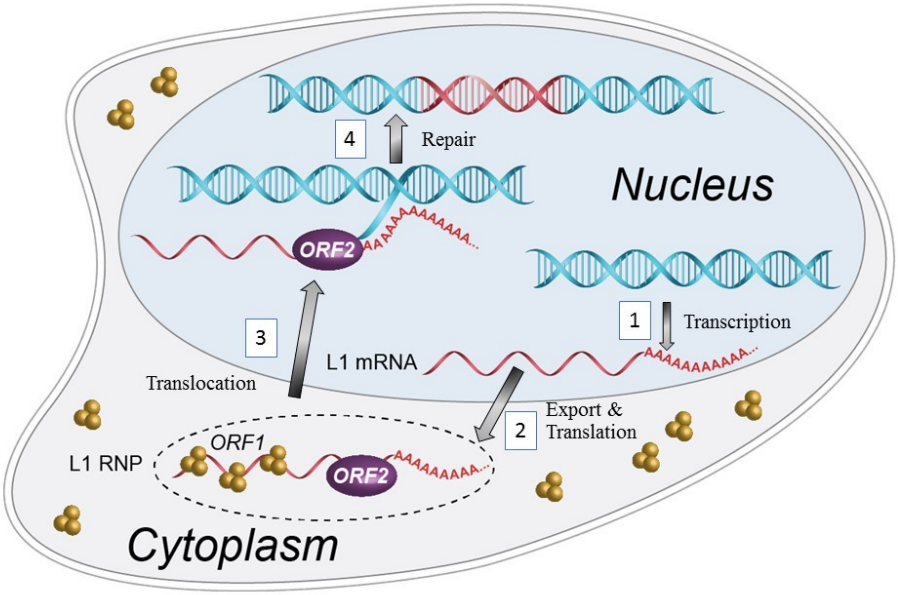


Figure 1.2: Life cycle of the LINE-1 (L1) element. (1) L1 RNA is transcribed from the DNA and processed like mRNA for export from the nucleus. (2) The RNA is exported from the nucleus into the cytoplasm where ORF1p and ORF2p are translated from the bicistronic RNA. (3) The proteins and the L1 RNA form a complex which is then translocated into the nucleus. The DNA is nicked by the endonuclease domain of ORF2p and then reverse transcription occurs to reintegrate the L1 into a new genomic location. (4) Finally, DNA repair occurs ensuring the newly inserted L1 element is present on both strands and the DNA molecule is intact.

Due to the activity and the necessity of its proteins for mobilization, the L1 is the only autonomous transposable element (TE) active in the human genome. For TEs like *Alu* and SVA or processed psuedogenes to mobilize, these RNAs must 'hijack' the L1 machinery and are then assimilated into new locations in the human genome. Although *Alu* elements are non-autonomous, they are very abundant in the human genome with as many as 1 million copies total. *Alu*s are much smaller than L1 at around 300 nucleotides and do have an internal pol III promoter to contribute to their transcription. Interestingly, *Alu* elements are notorious for causing non-allelic homologous recombination (NAHR) in the genome (46). *Alu* elements are derived from two 7SL RNAs which fused prior to the evolution of primates (47). L1 ORF2p has mobilized marked *Alu* elements *in vitro* using a cell culture assay (48) and there is ample evidence of L1's ability to mobilize *Alu* elements *in vivo* in the human genome. *Alu* elements also bear the hallmarks of TPRT with TSDs and poly(A) tails at the sites of insertion (49). L1 also mobilizes a hominid specific (50) family of elements in the human genome known as SINE-VNTR-*Alu* (SVA). SVA elements are generally 2,000 nucleotides in length and like L1 insertions and *Alu*s, have TSDs and poly(A) tails (49,51). Interestingly, SVA insertions vary

greatly in size because of variation in the VNTR repeat region (50) and the presence or absence of transductions (52–54). The elements primarily consist of four domains, in order from the 5' side: (i) a CT rich repeat, also referred to as a hexamer, with CCCTCT serving as the most common motif, (ii) a sequence which shares homology to antisense *Alu*-like fragments, (iii) a variable number of GC-rich tandem repeats (VNTR), and an ~490 nucleotide sequence derived from the envelope gene and right long terminal repeat (LTR) of an extinct HERV-K10 containing a canonical poly(A) signal. Using the cell culture assay, ORF2 has been shown to mobilize SVA *in vitro* as well; furthermore, evidence suggests that the CT hexamer and the *Alu* like fragment of the SVA have a synergistic effect on the rate of retrotransposition (55). In addition to the aforementioned non-autonomous transposable elements *Alu* and SVA, L1 machinery also mobilizes processed psuedogenes (30,31).

Retrotransposon contribution to genomic instability and disease

L1 sequences comprise approximately 17% of the genome; however, taking into account the mobilized *Alu*, SVA, and pseudogene insertions which occurred due to L1 protein expression, L1 is responsible for ~33% of the sequence of the genome (4). Most of the transposable elements in the human genome are no longer identifiable as TEs due to the age of their sequences and their sequence divergence is so great that they cannot be assigned to a single TE family (4,56). The majority of the L1 elements in the human genome are not actively transposing because only one subfamily of each type of elements is active at a time (56,57). Subfamilies are determined by differences in sequence content which occur as mutations accrue over evolutionary time in the respective elements. Occasionally there is overlap between L1 subfamilies, but the overlap is usually brief (58). The emergence of new active subfamilies of elements is largely due to a situation referred to as an "arms race", a part of the 'Red Queen' hypothesis (59). The arms race serves the purpose of evading the host defenses and the race is waged between both retrotransposons and the mechanisms for controlling their activity, such as APOBEC3 proteins (60–63). To limit the activity of potentially mutagenic TE insertion events, eukaryotic cells have acquired several defense mechanisms to derail the process of L1 mobilization at various stages. The fossils of older inactive elements are a testament to the fact that continued evolution of transposable elements occurs nearly constantly although the rate of retrotransposition has not been constant (57). One of the main tools utilized to limit retrotransposition is the methylation of retrotransposon promoters to restrict the transcription of the elements. Varied epigenetic modifiers are active in retrotransposition silencing, including the DNA methyltransferase-like protein Dnmt3L (64,65). Dnmt3L is essential for Dnmt3A mediated methylation of retrotransposons in primordial germ cells (64,65). In addition there is a

mechanism for controlling retrotransposon activity in the germline via a piRNA specific pathway which mediates genome-wide CpG methylation of TEs and restricts their activity (64,66–69). It has been demonstrated that L1 expression levels are inversely correlated with the methylation of the canonical promoter in the 5' UTR of the element (70–72), and numerous epigenetic modifiers contribute to establishing and maintaining the methylation status of L1 elements in the genome. To reinforce the suppression of TE activity, eukaryotic cells have also developed a Piwi-interacting RNA silencing pathway (A. Aravin et al., 2008; A. Aravin et al., 2006). Piwi-interacting RNAs, repeat associated small interfering RNAs, and microRNAs all act to degrade retrotransposon transcripts via RNA interference (74–79). RNA interference is yet another mechanism by which the host can control TE expression using repeat-associated small interfering RNAs and micro-RNAs to degrade TE transcripts (61,74–76,78–84). In addition to epigenetic and post-transcriptional regulation of L1, there are numerous host factor proteins which target the process by which L1s and other retrotransposons integrate into the genome. Oftentimes, these host factors are also used to control retroviral infection in the host. MOV10, a host factor, is a potential RNA helicase and has the ability to restrict L1 retrotransposition in cell culture (85–87). MOV10 restricts TEs by associating with the key RNA-induced silencing complex component AGO2 and the L1 ribonucleoprotein particle (RNP). After association with AGO2 and the RNP, MOV10 is theorized to degrade or block the translation of L1 RNA (88). The exonuclease Trex1 metabolizes reverse transcribed retrotransposon DNA to stunt the process of retrotransposition (89). In addition to Trex1 and MOV10, many studies have reported members of the APOBEC3 (A3) family of cytidine deaminases having a role in restricting the activity of L1 elements in cultured cells (60,62,63,90).

For TE families, especially L1, to continue to propagate in the human genome, elements must evolve to circumvent cellular host control mechanisms. In addition to the competition between the host and the TE, the TEs themselves must compete among other elements to retain their activity as well. The competition among L1 elements has been demonstrated *in vivo* in studies of rodent L1s (91,92) and further supported by evidence that L1 subfamilies seem to only coexist when the elements contain differing 5' UTR sequences (57). In the mouse genome there are three subfamilies of active L1 elements and all have sequence differences in their 5' UTR (93).

The currently active L1 elements, L1PA1 and L1Hs, are the products of a long succession of L1 element evolution. The active elements can be sub-classified based on certain 'diagnostic' nucleotides. Elements in the transcribed group a subfamilies are referred to as 'Ta' elements (94) appeared approximately 2-3 million years ago (95,96) and have "ACA" at positions 5924-5926 and a "G" at 6010 relative to the active $L1_{RP}$ element. Older and inactive elements instead have a "GAG" and an "A" at the same positions. A family which is slightly older, yet still active, known as the pre-Ta elements have the sequence "ACG" in place of "ACA". These diagnostic nucleotides are key factors for the selection process during the library preparation for this body of work which will be discussed in a subsequent chapter.

Only a subset of the preTa and Ta L1 elements are capable of transposing to new locations in the genome and there are several requirements which must be met. Because many L1s are 5' truncated upon insertion, a large number of the elements are unable to promote their own transcription because the 5' promoter is absent from the insertion site (97). The 5' truncations may be due to poor processivity during the reverse transcriptase reaction or potentially because of degradation of the L1 RNA after translation and prior to the reverse

transcription. During the insertion process, oftentimes the transcript is modified and results in frame shifts or other inactivating mutations in either of the ORFs which causes them to be potentially inactive. It has been determined that approximately 80 - 100 L1 elements are active in a diploid genome and are therefore able to mobilize themselves and other TEs in *trans* (98). Both copies of the L1 present in a diploid genome can generate new insertions (98). Additionally, there are approximately 2000-3000 *Alu* elements and less than 100 SVA elements capable of retrotransposition in the genome (99,100). There is allelic variability between L1 elements (101,102). In a mechanism not dissimilar to single nucleotide polymorphisms (SNPs) (103), SNPs in active L1 elements can change the activity up to 16-fold (102). A study comparing 3 active L1s across ~200 haploid genomes from six geographic regions resulted in 0% to 390% activity when compared to a reference (101). In this study comparing a trio of L1s, it was also noted that one new L1 allele (i.e. the same L1 with a different nucleotide sequence variant) existed for every 3-5 L1s sequenced in the study. Because the active elements are mobilizing to novel insertion sites in the genome, it is logical that individuals will differ with respect to the presence or absence of L1 insertions at loci throughout their genomes. These retrotransposons insertion polymorphisms (RIPs) segregate with populations in much the same way that SNPs do. Because many insertions derived from retrotransposons which are active occurred recently, they are polymorphic with regard to the presence or absence of the insertion in different human populations (95,96,104–107). In a study using fosmid end resequencing and mapping to identify 6,000 nucleotide and greater structural variants, 68 non-reference L1 RIPs were identified (108). Of the 68 RIPs identified, 37 were found to be “hot” or highly active when assayed in cell culture using the retrotransposition assay (25). In addition to the new RIPs discovered, the authors noted that each of the six individuals studied possessed 2/6 insertions

present in the reference genome that were classified as “hot” in previous work (98). In addition to the 2 aforementioned “hot” L1 elements, each individual possessed between 3 and 9 additional “hot” elements which were not in the reference genome (108). Altogether, these studies demonstrate L1 is active and mobilizing in the genome.

Spontaneous and inherited occurrences of disease causing mutations have been observed in greater than 100 cases, including diseases such as hemophilia, cancer, and diabetes (109–175) (see Table 1.1). Previously, it has been suggested that ~0.27% of human genetic disease is caused by TE insertions (176). An example of a somatic insertion causing disease would be a processed pseudogene which inserted into the CYBB gene and caused primary immunodeficiency(177). Yet another example of a somatic event causing disease occurred when SVA mediated deletions in the NF1 gene caused disease(178). A somatic L1 insertion caused Choroideremia in a patient when it inserted into the coding region of the gene(179). There are many mechanisms by which TEs could disrupt normal gene expression or affect genome structure. TEs can disrupt genomic sequences when they insert; however, they can also cause deletions and rearrangements in the genome (via 5’and 3’ transduction) (112,180). Transductions can occur from both non-reference and reference L1 elements and are a result of the weak poly(A) signal in the L1 element (158,181–183). Because of the weak poly (A) signal, RNA polymerase II reads through the L1 to the adjacent DNA following the 3’ end of the element. This process is estimated to occur in 15- 23% of all L1 mobilization events (158,181–183). An L1 mediated 3’ transduction of a novel noncoding gene into exon 67 of the dystrophin gene was observed; however, due to severe 5’ truncation of the element there was no recognizable L1 sequence present (184,185). When L1s carry regulatory sequences in the transduction, “exon shuffling” can also occur which can affect gene expression (186). Yet

another mechanism by which L1 insertions can cause aberrant gene expression is “gene breaking” (187). For gene breaking to occur and L1 must insert into an intron in the antisense orientation and then split the associated transcript into two parts through the combined effects of the L1 polyadenylation signal and the L1 antisense promoter (187). TEs can also provide an alternative promoter for a gene following their insertion into a new location. A TE which is fixed in the genome or even polymorphic in the population can acquire mutations which enable the sequence to create a cryptic splice site (188) or it can undergo deletions which facilitate branch site recognition and result in *Alu* exonization (189).

In addition to insertional mutagenesis, retrotransposons can mediate ectopic recombination through non-allelic homologous recombination (NAHR) and non-homologous end joining (NHEJ) in the genome (46,190–194). In fact, a frequently observed example of this process is also the most frequently observed translocation in the human genome where there is a recombination of two *Alu* sequences on chromosomes 11 and 22 respectively (195). Additionally, it has been demonstrated that *Alu* repeats are enriched in segmental duplication breakpoints (196) and countless examples of NAHR mediated by *Alu* elements have been found (192). Recently a broad analysis of pathogenic variants in Fanconi anemia genes found that up to 75% of FANCA deletions are *Alu*-*Alu* mediated, predominantly mediated by NAHR due to *Alu* Y elements (197). Occasionally the homologous sequences of L1 elements cause mis-alignment during meiosis and result in NAHR especially when elements are proximal to each other and in the same orientation (190–193). A 520 kb deletion containing four genes occurred due to an L1 associated non allelic recombination and caused Ellis von Creveld syndrome in a family (198). Other previous reports also noted a recombination between L1 elements flanking the PHKB gene (199) and a similar event occurred due to the same mechanism causing Alport

syndrome diffuse leiomyomatosis (173). More recently, a deletion in the factor IX gene between two highly homologous L1 sequences seems to have occurred due to non-allelic-homologous recombination between the two tandem L1s (200). SVA elements, through NAHR, are also responsible for disease-causing mutations occurring due to copy number changes with non-recurrent breakpoints (178). In a recent study by Vogt and colleagues, large NF1 deletions were studied and two of 17 deletions with non-recurrent breakpoints occurred with the concomitant insertion of SVA elements at the deletion breakpoints (178).

Many of the previously described disease causing events associated with the presence of retrotransposons in the human genome occur during a post-zygotic stage (178) or occur in the germline. Although diseases caused by insertion or aberrant recombination events have mostly been due to insertions prior to or during development, insertions occurring in somatic cells of diseased organisms have also been clearly exhibited. Although the somatic insertions appear to be occurring in various regions of the brain, a subset of normal tissues, and most epithelial cancers, it has yet to be determined to what extent the insertions are changing gene expression and potentially contributing to human disease. Somatic insertions in cancer will comprise the bulk of this thesis and will be further discussed in the next section (201).

Due to the various mechanisms through which repetitive elements can cause disease, the host has evolved many pathways to thwart the amplification of retrotransposons and thereby the potential mutations which come along with them. The defense mechanisms employed by the cell are diverse as they affect various aspects of the L1 life cycle. The previously discussed L1 control mechanisms display how the host uses multiple, if not redundant, mechanisms to control retrotransposon mobility and suggest when any of these mechanisms is not operating optimally L1 may be more active. In other words, a cell subject to aberrant expression of its protective

mechanisms may be particularly susceptible to L1 somatic insertions that are inherently mutagenic.

Element	Subfamily	Gene	Disease	Chr	Reference	Size (bp)	PolyA tail length (bp)
Alu	AluYb9	*ABCD1*	Adrenoleukodystrophy (ALD)	X	113	98	20
Alu	AluYa5a2	*ATP7A*	Menkes Disease	X	114	282	89
Alu	AluY	*BTK*	X-linked agammaglobulinemia (XLA)	X	111	N/A	N/A
Alu	AluY	*BTK*	X-linked agammaglobulinemia (XLA)	X	115	281	74
Alu	AluYb8	*CD40LG*	Hyper-immunoglobulin M syndrome (HIGM)	X	116	292	8
Alu	AluYa5	*CLCN5*	Dent's Disease	X	111,117	281	50
Alu	Alu	*CTRC*	Chronic pancreatitis	1	118	31	11
Alu	AluYb8	*FVIII*	Hemophilia A	X	119	290	47
Alu	AluYb9	*FVIII*	Hemophilia A	X	120	288	37
Alu	AluYb8	*FVIII*	Hemophilia A	X	121	FL	N/A
Alu	AluYa5a2	*FIX*	Hemophilia B	X	122	244	78
Alu	AluYa5a2	*FIX*	Hemophilia B	X	123	237	39
Alu	AluY	*FIX*	Hemophilia B	X	124	279	40
Alu	AluYc1	*GK*	Glycerol kinase deficiency (GKD)	X	125	241	74
Alu	AluYa5	*IL2RG*	X-linked (XSCID)	X	111	N/A	N/A
Alu	AluY	*CRB1*	Retinitis pigmentosa (RP)	1	126	244	70
Alu	Alu	*SERPINC1*	Type 1 autoimmune thyroid disease (ATD)	1	127	6	40
Alu	AluYa5	*ALMS1*	Alström syndrome	2	111	257	76
Alu	AluJ	*MSH2*	Lynch syndrome or hereditary nonpolyposis colorectal cancer (HNPCC)	2	128	85	40
Alu	N/A	*MSH2*	Hereditary Cancer	2	175	N/A	N/A
Alu	AluYa5	*ZFHX1B*	Mowat-Wilson syndrome	2	111	281	93
Alu	AluYb9	*BCHE*	Cholinesterase deficiency	3	129	289	38
Alu	AluYa5	*CASR*	Familial hypocalciuric hypercalcemia and neonatal severe hyperparathyroidism (FHH and NSHPT)	3	130	280	93
Alu	AluYb8	*HESX1*	Anterior Pituitary Aplasia	3	131	288	30
Alu	AluYb8	*OPA1*	Autosomal dominant optic atrophy (ADOA)	3	132	289	25
Alu	AluYa5	*MLVI2*	Associated with leukemia*	5	133	280	26
Alu	AluYb8	*APC*	Hereditary desmoid disease (HDD)*	5	134	278	40
Alu	N/A	*APC*	Hereditary cancer	5	175	N/A	N/A
Alu	AluYb9	*APC*	Familial adenomatous polyposis (FAP)	5	135	93	60
Alu	AluY	*MCC*	hepatocellular carcinoma	5	144	N/A	N/A
Alu	AluYb8	*MAK*	Retinitis pigmentosa (RP)	6	136	281	57
Alu	AluYa5	*NT5C3*	Chronic hemolytic leukemia (CHL)	7	137	281	36
Alu	AluY	*CFTR*	Cystic Fibrosis	7	138	46	57
Alu	AluYa5	*CFTR*	Cystic Fibrosis	7	138	281	56
Alu	AluYa5	*EYA1*	Brancio-oto-renal (BOR) syndrome	8	139	N/A	97
Alu	AluYb9	*LPL*	Lipoprotein disease (LPL) deficiency	8	111	150	60
Alu	AluYb5/8	*CHD7*	CHARGE syndrome	8	140	75	100
Alu	AluYa5	*POMT1*	Walker Walburg syndrome	9	141	290	53
Alu	AluYa5	*FGFR2*	Apert syndrome	10	142	283	69
Alu	AluYb8	*FGFR2*	Apert syndrome	10	142	288	47
Alu	AluYk13	*FGFR2*	Apert syndrome	10	143	214	12
Alu	AluYa5	*FAS*	Autoimmune lymphoproliferative syndrome (ALPS)	10	144	281	33
Alu	AluYc1	*SERPING1*	Hereditary form of angioedma (HAE)	11	145	285	42
Alu	AluYa5	*HMBS*	Acute intermittant prophyria (AIP)	11	146	279	39
Alu	N/A	*ATM*	Hereditay cancer	11	175	N/A	N/A
Alu	AluYa5	*GNPTAB*	Mucopolydosis Type II (ML II)	12	147	279	17
Alu	AluYc1	*BRCA2*	Breast Cancer	13	148	281	62
Alu	N/A	*BRCA2*	Breast Cancer	13	175	N/A	N/A
Alu	N/A	*BRCA2*	Breast Cancer	13	175	N/A	N/A
Alu	N/A	*BRCA2*	Breast Cancer	13	175	N/A	N/A
Alu	N/A	*BRCA2*	Breast Cancer	13	175	N/A	N/A
Alu	N/A	*BRCA2*	Breast Cancer	13	175	N/A	N/A
Alu	N/A	*BRCA2*	Breast Cancer	13	175	N/A	N/A
Alu	N/A	*BRCA2*	Breast Cancer	13	175	N/A	N/A
Alu	N/A	*BRCA2*	Breast Cancer	13	175	N/A	N/A
Alu	AluYa5	*BRCA2*	Breast Cancer	13	149	285	N/A
Alu	N/A	*PALB2*	Hereditary cancer	16	175	N/A	N/A
Alu	AluYb8	*PMM2*	Congenital disorders of glycosylation type Ia (CDG-Ia)	16	150	263	10
Alu	AluYc1	*BRCA1*	Breast and Ovarian Cancer, Familial	17	151	191	60
Alu	N/A	*BRCA1*	Hereditary Cancer	17	175	N/A	N/A
Alu	AluS	*BRCA1*	Breast Cancer	17	149	286	N/A

Element	Subfamily	Gene	Disease	Chr	Reference	Size (bp)	PolyA tail length (bp)
Alu	AluYa5	*NF1*	Neurofibromatosis type 1 (NF1, cancer)*	17	152	282	40
Alu	AluY	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	280	N/A
Alu	AluY	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	281	N/A
Alu	AluYa5	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	282	60
Alu	AluYa5	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	284	120
Alu	AluYa5	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	281	N/A
Alu	AluYa5	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	284	110
Alu	AluYa5	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	279	N/A
Alu	AluYa5	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	264	60-85
Alu	AluYb8	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	249	121
Alu	AluYb8	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	288	N/A
Alu	AluYb8	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	289	120
Alu	AluYb8	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	288	78-178
Alu	AluYb8	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	288	118
Alu	AluYb8	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	268	121
LINE-1	L1 Ta	*CYBB*	Chronic granulomatous disease (CGD)	X	154, 155	1722	101
LINE-1	L1 Ta	*CYBB*	Chronic granulomatous disease (CGD)	X	155	836	69
LINE-1	L1 Ta	*CHM*	Choroideremia	X	116	6,017	71
LINE-1	L1 Ta	*DMD*	Duchenne muscular dystrophy (DMD)	X	156	452	41
LINE-1	L1 Ta	*DMD*	Duchenne muscular dystrophy (DMD)	X	157	608	16
LINE-1	L1 Ta	*DMD*	Duchenne muscular dystrophy (DMD)	X	111	1400	38
LINE-1	L1 Ta	*DMD*	Duchenne muscular dystrophy (DMD)	X	158	530	73
LINE-1	N/A	*DMD*	Duchenne muscular dystrophy (DMD)	X	E Bakker & G van Omenn, pers.comm.	878	N/A
LINE-1	L1 Ta	*DMD*	Duchenne muscular dystrophy (DMD)	X	122, 123	212	118
LINE-1	L1 Ta	*FVIII*	Hemophilia A	X	109	3800	54
LINE-1	L1 preTa	*FVIII*	Hemophilia A	X	109	2300	77
LINE-1	L1 Ta	*FIX*	Hemophilia B	X	124	463	68
LINE-1	L1 Ta	*FIX*	Hemophilia B	X	159	163	125
LINE-1	L1 Ta	*RP2*	X linked retinitis pigmentosa(XLRP)	X	111	6000	64
LINE-1	L1 HS	*RPS6KA3*	Coffin-Lowry Syndrome	X	111	2800	N/A
LINE-1	N/A	*ABDH5*	Chanaric-Dorfman syndrome (CDS)	3	Sprecher pers. comm.	FL	N/A
LINE-1	N/A	*MLH1*	Hereditary Cancer	3	175	N/A	N/A
LINE-1	N/A	*MLH1*	Hereditary Cancer	3	175	N/A	N/A
LINE-1	L1 Ta	*APC*	Colon cancer	5	176	520	222
LINE-1	L1 Hs	*EYA1*	Branchio-oto-renal syndrome (BOR)	8	160	3756	None
LINE-1	L1 Ta	*ST18*	Hepatocellular carcinoma*	8	144	410	N/A
LINE-1	L1 Ta	*FKTN*	Fukuyama muscular dystrophy (FCMD)	9	161	1200	59
LINE-1	L1 Ta	*FKTN*	Fukuyama muscular dystrophy (FCMD)	9	161	3000	N/A
LINE-1	L1 Hs	*SETX*	Ataxia with oculomotor apraxia type 2 (AOA2)	9	162	1300	42
LINE-1	L1 Ta	*HBB*	β thalassemia	11	163	6000	107
LINE-1	L1 Hs	*PDHX*	PHHc deficiency	11	164	6086	67
LINE-1	L1 Ta	*SLCO1B3*	Rotor syndrome	12	165	6,100	N/A
LINE-1	L1 preTa	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	1800	N/A
LINE-1	L1 Ta	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	6,000	N/A
LINE-1	N/A	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	2200	N/A
LINE-1	L1 Ta	*PTEN*	Endometrial carcioma	10	149	112	N/A
SVA	F	*FVIX*	Hemophilia B	X	166	2524	28
SVA	F_1	*SUZ1P*	Neurofibromatosis type 1 (NF1, cancer)*	17	115	1700	23
SVA	F	*SUZ1P*	Neurofibromatosis type 1 (NF1, cancer)*	17	115	1300	40
SVA	F	*BTK*	X linked agammaglobulinemia (XLA)	X	115	251	92
SVA	F	*TAF1*	X linked dystonia-parkinonism syndrome (XDP)	X	167	2627	62
SVA	E	*LDRAP1*	Autosomal recessive hypercholesterolaemia (ARH)	1	168	2600	57
SVA	E	*SPTA1*	Hereditary Elliptoytosis and Hereditary Pyropoikilocytosis (HE and HPP)	1	111	632	50
SVA	F	*HLA-A*	Leukemia	6	169	2000	45
SVA	F	*PMS2*	Lynch syndrome	7	202	2200	N/A
SVA	E	*FKTN*	Fukuyama muscular dystrophy (FCMD)	9	170,171	3023	32
SVA	E	*PNPLA2*	Neuroal lipid storage disease with myopathy (NLSDM)	11	172	1800	44
pA	N/A	*COL4A6*	Alport syndrome	X	173	N/A	70
pA	N/A	*AGA*	Aspartylglucosaminuria (AGU)	4	111	N/A	37
pA	N/A	*BRCA2*	breast cancer	13	174	N/A	35
pA	N/A	*NF1*	Neurofibromatosis type 1 (NF1, cancer)	17	153	N/A	120
PP**	*TMF1*	*CYBB*	Chronic granulomatous disease (CGD)	X	114	5800	100

Table 1.1: Disease-causing retrotransposon insertions. This table details an up to date list of all known disease causing retrotransposon mediated mutations. The diseases with a * following their name indicate insertions which are known to be somatic. In this table, the type of element and the location into which it inserted, the disease caused by the insertion, the size of the insertion, and the length of the poly(A) tail are listed. The ** indicates a processed pseudogene.

L1 and Cancer

It follows logically that genetic instability caused by retrotransposition activity would be elevated in diseases where normal cellular check points during proliferation and DNA replication are absent, such as cancer. Indeed, many cancers have shown high L1 expression and a high occurrence of somatic insertions in patients evaluated thus far (202–210). Although L1 activity in cancer, especially epithelial cancers, is prevalent, it is still unclear how much somatic retrotransposon insertions are contributing to oncogenesis. Furthermore, it is still unclear if the relationship between cancer and retrotransposition in epithelial cancers is due to cancer activating the process of retrotransposition or due to retrotransposition causing somatic mutations which contribute to tumor formation. Cancer is by no means a simple disease, in fact, it encompasses a wide-ranging group of more than 200 diseases that involve uninhibited proliferation of cells leading to tumor formation, in addition to several additional common features (211,212). Epidemiological studies on twins suggest that environment plays a much clearer role in the process of tumorigenesis than genetics (213,214). For example, Sorenson and colleagues found that when an adoptive parent died as a result of cancer before the age of 50, the rate of mortality due to cancer for the adoptees increased (214). These findings suggest that a strong enough environmental mutagen will have a potent effect on individuals who live in the same environment despite differences in genetic background. Furthermore, another study found that the overwhelming contributor to cancer development in twins was the environment (213). The authors found that even when they considered cancer for which there was statistically significant evidence of cancer heritability, most twin pairs were discordant for presence of the cancer (213). Environmental factors contribute to sporadic cancer occurrence as much as 58-82%, as compared to the highest known genetic contribution to cancers, colorectal, breast, and

prostate cancers which is 27-42% (213). Differentiating between mutations which contribute to cancer development, referred to as drivers, and those which accumulate due to the dysregulation of DNA replication and repair pathways, dubbed passengers, is a continuous challenge in the study of cancer genetics. The apparent dysregulation of L1 elements in cancer is only one of many sources of genetic aberrations that frequently contribute to cancer development. However, L1 elements and other retrotransposons have a large size effect upon insertion and due to their structure have multiple ways in which their newly acquired presence can disrupt gene expression and regulation. It is also telling that L1 mobilization has been observed in many different tumors (Doucet-O'Hare et al., 2015; Ewing et al., 2015; Helman et al., 2014; Lee et al., 2012; Rodić et al., 2015; Shukla et al., 2013; Solyom, Ewing, Rahrmann, et al., 2012b; Tubio et al., 2014), cancer cell lines (25,215,216), and during development (217–219). Due to the potentially substantial effect of an L1 insertion and the predominantly deleterious effects on host gene expression observed thus far (205,220,221), L1 insertions may be more prone than other types of mutations to have an impact on tumorigenesis.

There are many carcinogenic environmental factors (222) which have an impact on retrotransposon activity in cultured cells (223). Benzoprenes, for example, are a risk factor for lung cancer, colorectal cancer, and breast cancer (224–226) and have been shown to increase L1 retrotransposition in HeLa cells (227). Exposure to certain metals such as cadmium, chromium VI, and nickel are risk factors for lung and breast cancer (228,229) and interestingly, nickel has been shown to induce a higher rate of L1 retrotransposition (230). Another feature of many tumors is a higher level of free radicals involved in oxidative stress (231) and oxidative stress has also been demonstrated to affect L1 activity (232). Furthermore, oxidative stress and DNA damage frequently occur as a result of cellular senescence and can also increase both

retrotransposition rates and chromosomal instability thereby potentially contributing to somatic mosaicism and cancer development (233–235). It is certainly plausible that many more environmental factors which contribute to cancer development also activate L1 retrotransposition and thereby increase the probability of L1 generating an insertion which affects an oncogenic locus and contributes to tumorigenesis (236).

When a cancer genome is sequenced tens or hundreds of thousands of single nucleotide variants, insertions, deletions, translocations, rearrangements, and other mutations may be found. In order to understand the role L1 mobilization plays in tumorigenesis, it is necessary to separate the winnow from the chafe, determine whether any somatic L1 insertions are present in the tumor, and absent from the normal tissues. To determine whether or not somatic insertions are contributing to tumor development, the insertions must be mapped in individuals with relevant disease. In 1992, Miki et al. mapped a somatic L1 insertion in a colorectal tumor in an exon of the APC gene (237). The somatic insertion was confirmed with Southern blot and because *APC* is the primary tumor suppressor gene in colorectal cancer and causes familial adenomatous polyposis (238,239) it is reasonable to conclude that the somatic L1 insertion, which was found to be absent from normal colon, was sufficient to drive oncogenesis (236). Although the preliminary discovery of an L1 insertion contributing to cancer occurred in the early 90s, it was two decades before researchers returned to the topic to investigate the role of L1 in carcinogenesis. To date, only one other definitive somatic insertion has been found in the exon of a tumor suppressor gene. An insertion into an exon of the PTEN gene was discovered with whole genome and whole exome sequencing by Helman and colleagues in 2014 (210). High-throughput next-generation sequencing enabled researchers to examine the genomes of more individuals at one time and compare those genomes between the cancer and normal samples in

addition to comparing individuals' genome differences. Due to the new technology available, many methods were subsequently specifically developed for assessing L1 activity in the genome, for detailed reviews see (240,241). Prior to a paper from Iskow and colleagues, several groups were able to successfully identify novel L1 insertions; however, they used assays which were inherently low-throughput and which had high false positive rates (105,242,243). The initial high-throughput method utilized for the discovery of somatic insertions, termed 'Transposon-seq' utilized digested genomic DNA using restriction enzymes which recognize the 3' end of the L1 and *Alu* elements (244). The authors linked adapters to the resulting fragments and amplified them with PCR to create retrotransposon specific libraries (244). In the initial efforts of the study, 38 ethnically diverse humans and 8 ATCC cell lines derived from human tumors were utilized to create libraries (244). Approximately 4600 library fragments were cloned and sequenced with ABI capillary sequencing yielding 152 putative novel L1 insertion polymorphisms (244). In order to ensure a low false positive rate, the authors applied specialized informatics to filter the datasets and identified high probability L1Ta insertion candidates (244). The PCR validation rate for the insertions was 97% with approximately a third of the insertions possessing a minor allele frequency (MAF) equal to or below 5% (244). Additionally, 47 'rare' insertions were found in very few individuals and 9 of them were only found in one cell line evaluated (244). One of the nine rare insertions in only one cell line was deemed as a somatic insertion due to its presence in the tumor cell line and its absence from the normal cell line (244). After finding the somatic insertion, the authors implemented their technique in a high-throughput fashion by acquiring 20 non-small cell lung cancers with matched normal tissues. Previous work in the mouse brain (245) and in human neural stem cells (Coufal, Garcia-Perez, Peng, Yeo, Mu, Lovci, Morell, Oâ Shea, et al., 2009) suggested that L1 activity in

the brain was highly active. Two types of brain tumors were also evaluated in the study including glioblastoma and medulloblastoma (244) with 5 cases of each condition along with matched blood leukocyte controls (244). The high-throughput version of 'Transposon-seq' utilized barcoding sequences to assign a given sequence to specific samples within the sample pool sequenced with 454 pyrosequencing (244). Following sequencing analysis, 1389 distinct L1 insertions were detected in the 30 samples assessed. After all the novel insertion candidates were compared to the human reference genome and to dbRIP (107), 650 putative novel L1 insertions remained, and 45% of them had MAFs less than or equal to 5%. Of all the individuals evaluated, 93% of the genomes had at least one rare L1 insertion present in only a single human in the study. After screening the low frequency alleles with PCR assays, the authors found there were 9 tumor specific somatic L1 insertions present in their lung cancer cohort. Surprisingly, no somatic insertions were confirmed in the brain tumors evaluated. In six of the 20 lung tumors studied, somatic tumor-only insertions were confirmed. Lastly, the authors confirmed hypomethylation of many potentially active polymorphic L1 elements in the genomes of affected patients (244). The hypomethylation present in the affected individuals suggests that one mechanism of L1 escape from host control in cancer is due to changes in methylation due to mutations in tumor suppressor genes.

Several years later, Lee et al used a computational method, 'Tea' for transposable element analyzer, to analyze whole genome paired end sequencing data from tumors and matching blood samples (Lee et al. 2012). In this study, the authors performed a single nucleotide resolution analysis of retrotransposons in 43 high coverage whole-genome sequencing data sets from five cancer types (Lee et al. 2012). The study samples consisted of colorectal tumors, ovarian tumors, prostate tumors, blood cancer, and brain cancer (Lee et al. 2012). The

authors identified 194 high-confidence putative somatic retrotransposon insertions in the samples of epithelial origin only, e.g. ovarian, prostate, and colorectal tumors (Lee et al. 2012). Of the 194 high-confidence putative insertions, 183 of them were purported to be L1s, 10 Alu elements, and 1 endogenous retrovirus (ERV) (Lee et al. 2012). It was later determined that the putative ERV insertion was likely caused by a microhomology-mediated break-induced repair mechanism (247). With regard to the PCR and capillary sequencing validation of the predicted somatic insertions, 25/26 insertions were validated in colorectal cancer and 13/13 insertions validated in ovarian cancer with an overall rate of 97% validation (Lee et al. 2012). Finally, the authors noted that somatic and germline L1 insertion sites differed in genomic distribution as well as epigenetic characteristics. When comparing germline insertions to somatic insertions, germline events are depleted from genes significantly, likely due to strong negative selection on the events (248). The authors assert that the retrotransposon insertions seem to provide a selective advantage in certain individuals and that the insertions occurred in genes commonly mutated in cancers and substantially disrupted their expression (Lee et al. 2012).

In a publication from our own group in the same year, two high-throughput sequencing techniques which enrich for retrotransposons in different ways were utilized. L1-seq, developed by Adam Ewing (249), utilizes a hemi-specific PCR based library construction method to enrich for the young, active subfamily of L1s in the genome. RC-seq (version 1), developed by the Faulkner lab (220), uses probes designed to bind the 5' and 3' ends of L1 and SVA elements and probes tiled across the full length of an *Alu*. The probes are tiled on an array and the sheared genomic DNA is applied to the array as the relevant sequences bind. This DNA later has adapters ligated to it and is minimally amplified with PCR using only 8 cycles (220). Using L1-seq on two cohorts of 16 total colorectal cancer patients with matched tumor and normal tissues,

26/40 and 37/51 high stringency somatic insertions were identified and validated respectively. An additional 9 out of 16 lower stringency insertions with lower read-counts and map scores were identified and validated between both cohorts as well. In total, L1-seq resulted in the 3' validation of 69/107 putative tumor-specific somatic insertions and both 5' and 3' validation of 35 of said insertions. As is typical of both previous and follow-up studies, one tumor had 17 insertions present while 3 others had no insertions. Most of the insertions identified had severe 5' truncation and averaged about 1kb in size. Five of the 16 colorectal cancer patient samples were barcoded, pooled, and analyzed by shallow, multiplexed RC-seq. Using RC-seq, 8 L1, 83 *Alu*, and 5 SVA somatic insertions different from those identified with L1-seq were predicted. Only one of the L1 insertions predicted was confirmed to be truly tumor-specific, and 11 high-confidence predicted L1 insertions identified by L1-seq were also identified with RC-seq. Of the remaining putative insertions, 6/8 L1s, 30/57 *Alu* elements, and 6/11 SVA elements were validated in both tumor and paired normal tissue.

A year later, Faulkner and colleagues published an updated version of RC-seq which was utilized to analyze retrotransposon activity in 19 hepatocellular carcinomas (HCC) (205). The HCC cases consisted of fresh frozen tissue from patients positive for HBV or HCV and matched normal tissue. In the new version of RC-seq, a liquid phase capture was utilized to increase the number of probes available for binding to increase efficiency; furthermore the sequences of the probes used were also refined and edited to be a more effective pool for binding active elements. The optimized version of RC-seq produced a 4-fold increase in reads which aligned to non-reference genome L1s per library sequenced. Twelve out of 17 potential somatic insertions were validated in tumor only with PCR and sequencing confirmed the L1 is active in HCC. No SVA or Alu element somatic insertions were confirmed in any of the patients; however, a single L1

insertion was confirmed in normal liver and was found to absent from the corresponding tumor. The insertion into normal liver is surprising because it had been previously assumed that retrotransposition was not an active process in somatic tissues with the exception of the brain (219–221,250). If somatic retrotransposition happens in the normal tissues of some individuals, it is possible that in those individuals it could cause mutations which lead to disease like cancer development. Interestingly, the authors noted three different germline insertions into the *MCC* gene, mutated in colorectal cancers, in three individuals with HCC. The germline insertions coincided with a strong inhibition of MCC as confirmed with immunoblot and qRT PCR. Although this study did not definitively address whether or not somatic insertions contribute to tumorigenesis, it did present evidence that in some individuals inherited polymorphic L1 insertions may play a role. It seems plausible to assume that if a germline insertion can cause such a reaction, then so too can a somatic insertion.

In 2014, yet another pipeline emerged for analyzing whole genome sequencing data from 200 tumor samples and their matched normal counterparts (210). The following cancers were analyzed in the study: lung adenocarcinoma, lung squamous cell carcinoma, ovarian carcinoma, rectal adenocarcinoma, colon adenocarcinoma, kidney clear-cell carcinoma, uterine corpus endometrioid carcinoma, head and neck squamous cell carcinoma, breast carcinoma, acute myeloid leukemia, and glioblastoma multiforme (210). The study identified 7,724 unique, non-reference germline insertion sites and approximately 65% of them are known retrotransposon insertion polymorphisms (RIPs) previously identified in other studies (108,207,244,249,251–253). In total 810 putative retrotransposon insertions were predicted in the cancer samples and absent from normal samples. The candidate insertions exhibited the hallmarks of TPRT including target site duplications averaging 15 nucleotides in addition to a canonical

endonuclease motif (34,254). However, forty-seven putative somatic retrotransposition events were selected for experimental validation. The 47 tested insertions were predicted across 21 individuals and four tumor types. Thirty-nine of the insertions (83%) were validated as tumor specific by PCR and sequencing of either the 5' or the 3' end. For 32 of the 47 insertions, evidence was present for both the 3' and 5' ends. Two of the putative somatic insertions were found to be germline, present in both tumor and normal, after the validation attempt. Six of the 47 putative somatic insertions were not amplified in either the normal or the tumor samples. Not unlike previous similar studies, it was noted that 97% of the L1 somatic insertions are in the L1Hs subfamily. After considering which cancers exhibited L1 activity among their samples the authors noted that cancers of epithelial origin were the only ones which had detectable somatic retrotransposition events. Historically, nearly all cancers found to possess retrotransposon activity, in the form of newly acquired somatic insertions unique to the tumor, have been epithelial cancers. Interestingly, the authors also observed several 3' transduction events from different regions of the genome in a single patient. The 3' transductions are evidence that at least three different source L1 elements contributed to the somatic insertions in the cancer. In contrast to this finding, the authors also noted a patient in which a single L1 element caused at least 4 events, detected due to their 3' transductions, into different areas of the genome. These findings seem to suggest two models for somatic retrotransposition activity in cancer. In some patients, a single hyperactive source element may insert itself into multiple genomic locations in the same tumor. In others, there may be several active source elements which contribute the somatic insertions present in the sample. It is also possible that both of these situations happen simultaneously in the same individual as well.

Later in 2014, a paper analyzed whole-genome sequencing data on 290 tumor and matched normal pairs consisting of 210 primary tumors, 52 metastatic tumors, and 28 cancer cell lines with matched normal cell lines (202). The samples were obtained from 244 patients across 12 cancer types including: bladder, bone, breast, colon, head and neck, lung, pancreatic, prostate, and renal cancer as well as mesothelioma, melanoma, and glioma (202). The algorithm used to analyze the sequencing data, "TraFiC', identified 2,756 putative L1 retrotransposition events including both 'solo' L1 events and 3' transductions. PCR validation was attempted on 308 putative insertions and 259 insertions were confirmed with PCR and capillary sequencing (202). The authors also observed a single patient with 22 somatic 3' transduction events from a hyper-active L1 which mobilized many times in the same cancer (202). The average insertion length was approximately 1 kb for insertions lacking a 5' inversion, the TSDs averaged between 10 and 20 base-pairs, and 3' transductions occurred in one fourth of the cancer genomes evaluated (202). Due to the abundance of 3' transductions in many of the samples, the authors were able to conclude that few loci were driving the 3' transductions in cancer (202).

Recently, another paper analyzing whole-genome paired-end sequencing was published in which the authors studied 43 cases of esophageal adenocarcinoma (255). The authors predicted an average of 16 insertions per tumor and a range of 0 to 153 insertions among the patients studied (255). One fifth of the L1 insertions found was predicted to have 5' inversions and there were 9 insertions identified with 3' transductions. The authors also attempted to correlate p53 loss with L1 activity by evaluating p53 mutations in all of the patients. The authors observed a p53 mutation present in over 88% of patients with esophageal adenocarcinoma patients studied; furthermore, two of the 5 cases where no p53 mutation were present were cases with no insertion (255). In addition, it is noted that it is not possible to conclude that the L1s are

only active in the tumor as the technique is not sensitive enough to detect potential events in the non-cancer cells due to their potentially highly polyclonal nature (255). A serious deficit in this study was the lack of matched normal tissues for nearly all patients. Without matched normal tissue, it is impossible to know whether putative somatic insertions detected in the cancer are truly somatic or if they occurred in early development and are present throughout the tissue of interest.

Using a technique dubbed 'Tip-seq', Rodic and colleagues studied 20 cases of pancreatic ductal adenocarcinoma (PDAC) to detect somatic L1 insertions present in the cancer and absent from normal pancreatic tissue (208). The authors had previously described L1 protein ORF1 expression in up to 89% of PDAC patients (256). Tip-seq is a PCR based L1 enrichment library preparation technique and it detected 268 somatic L1 insertions in the tumors of 18 patients evaluated which were absent from matched normal tissue (208). A range of 0 to 65 insertions was detected in the patients and an average of 15 insertions per case was calculated (208). There were 15 metastases which were evaluated with Tip-seq as well from 15 different patients and 242 insertions were detected in these samples (208). In 13 of the cases where both a metastasis and a primary tumor from a patient were shared, 45 insertions were confirmed by PCR and capillary sequencing to be present in both tissues and absent from the normal tissue (208). The expression of ORF1p in the samples subjected to Tip-seq correlated with the number of somatically acquired insertions per sample (208). The authors reported 81% of tested insertions validated with both PCR and capillary sequencing with all insertions being 5' truncated and an average size of approximately 1 kb (208). Finally, the authors noted two 3' transductions among the validated insertions in the study (208).

Ewing et al. published a study looking at multiple types of cancer including 4 colorectal cancer patients with matched colonic polyps and normal colon, 7 patients with pancreatic ductal adenocarcinoma with matched normal, 7 patients with gastric cancer and matched normal tissue, and 8 testicular germ cell tumors with matched blood (204). For 8 of the aforementioned cases, metastatic tissues were available and evaluated as well (204). Following L1-seq (249) and subsequent computational analysis 104 somatic heterozygous L1Hs insertions were validated by PCR and Sanger sequencing in the 18 gastrointestinal cancers and 1 insertion was validated in a single patient with a testicular germ cell tumor (204). However, the most interesting finding in this article was insertions occurring in the polyps which precede the cancer development (204). This pattern suggests that L1 is active in tissue before the cancer develops and certainly makes it seem more likely that L1 could contribute to the process of tumorigenesis.

All of the cancer studies performed to date have strongly established the hyper-activity of L1 in epithelial cancer; furthermore, many of the studies have established similar patterns with regard to retrotransposition. Several of the studies noted an average insertion size of approximately 1 kb likely due to the dramatic 5' truncations which most of the validated insertions possess(202,206,208). Thus far, nearly all the papers have reported target-site duplications in the same size range, approximately 10-20 nucleotides on average, and approximately 20% of the insertions detected in cancer have 5' inversions.

Although the activity of L1 elements in cancer has been firmly established and the events seem to adhere to most of the hallmarks of the process, it is still uncertain to what degree these elements play a role in carcinogenesis. Anything short of a glaringly obvious insertion disrupting a known tumor suppressor or activating an oncogene is a difficult sell to the scientific community as a cause or contributor to cancer. Furthermore, there is the possibility that the

dysregulation of normal cellular processes in cancer may simply be enabling L1 activity due to differences in methylation or the under expression of host genes which normally suppress retrotransposons activity. The evidence contrasting the simple activation of elements due to cancer development is the confirmed somatic insertions in not only the precursor conditions to cancer, but also in normal tissues. Observations of validated somatic insertions in tissues which are the precursor to cancer were made in the recent publication by Ewing and colleagues and have also been observed in work which will later be discussed thoroughly in this dissertation. Although there is mounting evidence of somatic retrotransposition occurring in normal tissues, it has not been definitively shown that this activity leads to cancer. Like any other potential mutagen, retrotransposition likely leads to disease a certain percentage of the time regardless of the disease type. However, when retrotransposons are hyper-active in a tissue, like in cancer or potentially precancerous conditions, it may be more likely to be the cause of a mutation which leads down the path to cancer development. In the future, to distinguish between retrotransposons as passengers verses drivers, single cell sequencing and the acquisition of a large cohort of patients will likely lead to an answer.

Areas of Study

The chapters which follow describe the utilization of an established technique to detect somatic L1 insertions occurring in cancer, in addition to the optimization of ORF1p detection via IHC in multiple types of cancer and normal tissues. This thesis will cover two distinct projects on esophageal adenocarcinoma and esophageal squamous cell carcinoma respectively, which characterize the activity of endogenous retroelements in both pre-cancerous conditions and cancer.

In Chapter 2, I describe the first of the two projects in which I utilized L1-seq (249) to evaluate the level of retrotransposition occurring in patients with Barrett's esophagus and esophageal adenocarcinoma. The subjects in the study consisted of 5 patients with benign (non-progressive) Barrett's esophagus, 5 patients with Barrett's esophagus which progressed to esophageal adenocarcinoma, and10 additional esophageal adenocarcinoma patients (206). Not only did I evaluate the insertions present in the tissues of the patients, but I also looked at ORF1p expression in many of the normal and cancer tissues (206). Interestingly, I observed mild ORF1p expression in many of the normal esophageal tissues evaluated from patients; however, the ORF1p expression was higher in the cancer samples. Furthermore, I observed sub-clonal insertions which were present in only a few cells in the normal tissue or the pre-cancerous tissue and then were expanded in the resulting lesion(206). This work was recently published in the Proccedings of the National Academy of Sciences (206).

In Chapter 3, I describe my second project, using L1-seq (249) to examine the L1 retrotransposon activity and ORF1p expression in nine patients with esophageal squamous cell carcinoma (Doucet-O'Hare et al., in prep). In these samples, I was able to observe more than 74 retrotransposition events where 12 events appeared to be sub-clonal in normal esophagus and

clonal in esophageal squamous cell carcinoma (Doucet-O'Hare et al., in prep). These data will be compiled into a publication for Oncogene and submitted in the coming weeks.

In Chapter 4, I will outline future directions for the areas of study which are discussed herein. The activity of L1 retrotransposons in cancer genomes is fairly well characterized; however, its potential contribution to tumorigenesis is poorly understood.

CHAPTER 2:

LINE-1 Expression and Retrotransposition in Normal Esophagus, Barrett's Esophagus, and Esophageal Carcinoma

Abstract

Barrett's esophagus (BE) is a common disease in which the lining of the esophagus transitions from stratified squamous epithelium to metaplastic columnar epithelium that predisposes individuals to developing esophageal adenocarcinoma (EAC). We hypothesized BE provides a unique environment for increased L1 retrotransposition. To this end, we evaluated 5 patients with benign BE, 5 patients with BE and concomitant esophageal adenocarcinoma (EAC), and 10 additional patients with EAC to determine L1 activity in this progressive disease. After L1-seq, we confirmed 118 somatic insertions by PCR in 10 of 20 individuals. We observed clonal amplification of several insertions which appeared to originate in normal esophagus (NE) or BE and were later clonally expanded in BE or in EAC. Additionally, we observed evidence of clonality within the EAC cases: specifically, 22 of 25 EAC-only insertions were present identically in distinct regions available from the same tumor suggesting that these insertions occurred in the founding tumor cell of these lesions. L1 proteins must be expressed for retrotransposition to occur; therefore, we evaluated the expression of ORF1p, a protein encoded by L1, in 8 of the EAC cases for which formalin-fixed paraffin embedded tissue was available. With immunohistochemistry, we detected ORF1p in all tumors evaluated. Interestingly, we also observed dim ORF1p immunoreactivity in histologically normal esophagus of all patients. In summary, our data show that somatic retrotransposition occurs early in many patients with BE and EAC, and indicate that early events occurring even in histologically normal esophageal cells may be clonally expanded in esophageal adenocarcinogenesis.

Introduction

Gastroesophageal reflux disease (GERD) affects a large proportion of Western populations and represents a significant healthcare burden partially due to its frequent evolution into Barrett's esophagus (BE) (257). BE was first described by Norman Barrett in 1950 (258) and is a common disease in which the lining of the esophagus transitions from stratified squamous epithelial cells to a cancer predisposing metaplastic columnar epithelium (258). The transdifferentiation increases cellular resistance to the low pH from the acid entering the esophagus through the sphincter separating it from the stomach (259). BE occurs in 8% to 20% of patients with GERD or about 3-8% of the total population (259). Furthermore, recent studies suggest that another 3-8% of the general population may have BE without symptoms (259).

The risk of a patient with BE developing the advanced premalignant lesion, high-grade dysplasia, or frank esophageal adenocarcinoma (EAC) is 0.5% per year; however, the five-year survival rate from EAC is only 13%-16% (259). Moreover, although the risk of malignancy is low, an EAC diagnosis is not usually made until the late stages of the disease when the illness is nearly incurable (259,260). Early diagnosis of dysplasia and EAC can be accomplished in patients with BE by screening endoscopies with biopsies performed at regular intervals determined by the physician (260). Due to the availability of tissue from these biopsies, detecting how the disease progresses and tracking clonal populations throughout disease progression has provided valuable insights into early cancer development (261,262).

Various types of mutations can be detected and subsequently monitored by biopsy to determine which clonal population of cells progresses to cancer. One source of mutation in epithelial cancer is retrotransposition (201–203,205,210). Retrotransposons compose approximately 45% of the human genome (111) and are mobilized via an RNA intermediate to

new genomic locations. The Long-INterspersed Element 1 (LINE-1 or L1) is the only autonomous retrotransposon that encodes the proteins necessary for mobilization and reinsertion into the genome. The two proteins encoded by L1 are responsible for the mobilization of other types of retrotransposons, *Alu* and SVA, as well as processed pseudogenes(31,48). Aside from contributing to genomic variation, retrotransposons can also have functional impact by inserting into transcription factor binding sites, donor and acceptor sites involved in mRNA splicing, enhancer sites, or protein coding regions of genes. To date, there are over 100 known retrotransposon insertions that caused single-gene diseases (110,111,193,263,264).

L1 mobilization in epithelial cancer has been observed by many groups at both the protein and DNA level. Interestingly, each individual has a different complement of 80-100 potentially active L1 elements in their genome which partially explains the large variation of somatic insertions detected in previous studies (201–203,205,210,249). Although it is evident that L1 is active and expressed in many cancer types, this activity has not been robustly evaluated in pre-cancerous lesions such as BE. New somatic insertions of L1 in BE could be used to track clonal progression of disease.

We hypothesized that the alterations in the esophageal lining as it undergoes cellular transdifferentiation present a permissive environment for retrotransposition. To test this hypothesis we evaluated individuals with BE who progressed to EAC, as well as those with non-progressive benign BE. We determined the occurrence of retrotransposition in these patients using L1-seq, a high-throughput L1-targeted sequencing method (20), and validated 118 somatic insertions in 10 of the 20 patients evaluated (Fig 2.1). Substantial levels of L1 protein expression were also detected in the EAC with immunohistochemistry; moreover, the protein was detected in the normal esophageal (NE) tissue of all patients tested. We conclude that this high

prevalence of L1 activity and insertions in BE and EAC, taken together with previous findings in other epithelial cancers, suggests a strong link between cancer and L1 activity. However, it is uncertain to what extent the dysregulation of normal cellular processes is contributing to L1 activation in cancer, as well as whether these somatic insertions are contributing to carcinogenesis in some individuals.

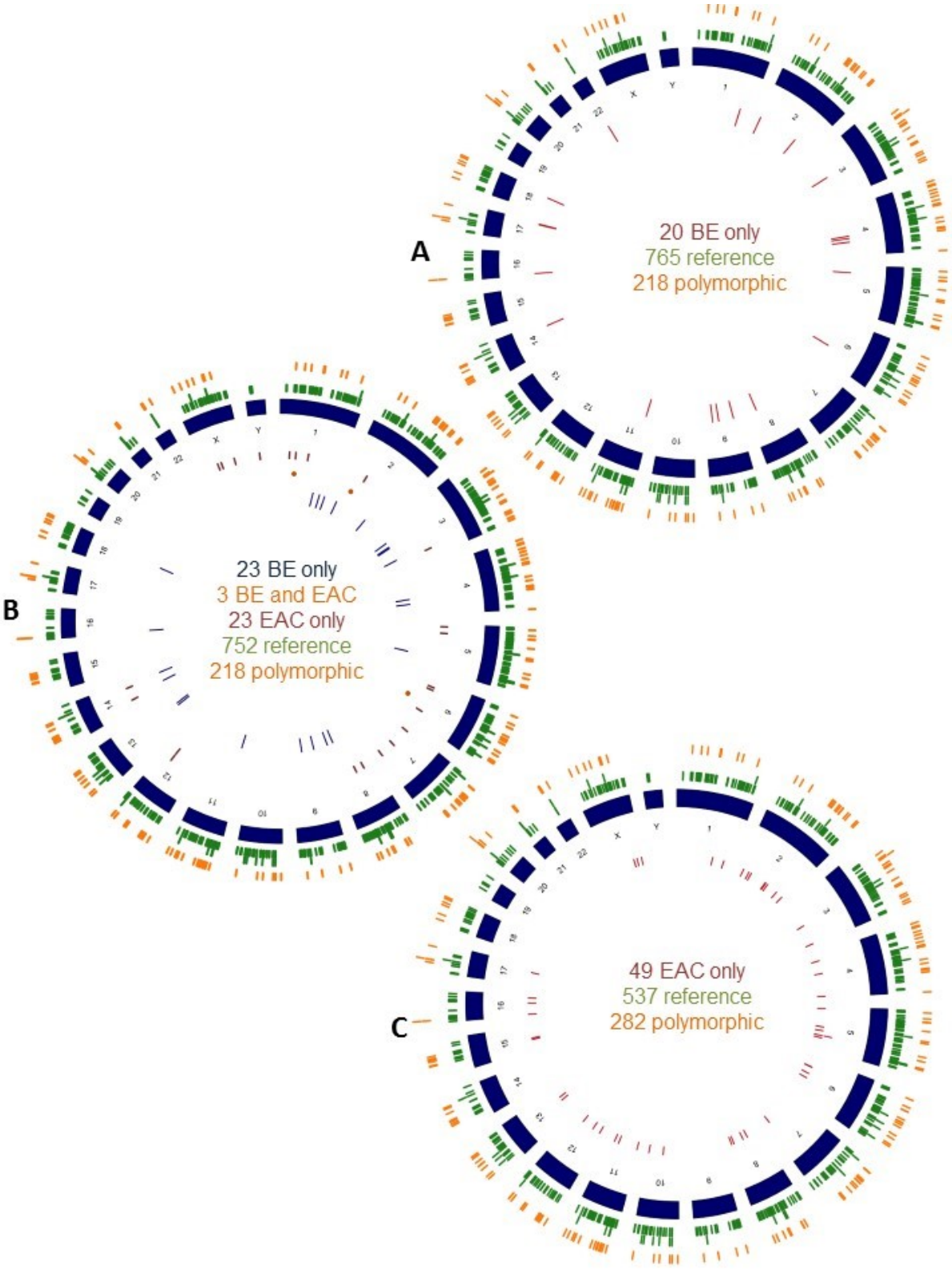

A
20 BE only
765 reference
218 polymorphic
B
23 BE only
3 BE and EAC
23 EAC only
752 reference
218 polymorphic
C
49 EAC only
537 reference
282 polymorphic

Figure 2.1: Distribution of reference, non-reference, and validated L1 somatic insertions in NE, BE, and EAC. A) Circos diagram mapping the distribution throughout the human genome of 20 validated high-stringency insertions in BE only (inner red circle), the 765 low stringency reference (outer green circle) and 218 polymorphic insertions (orange circle) detected with L1-seq. Group contains 5 individuals. B) Circos diagram mapping the distribution of the predicted 23 high stringency insertions in BE only (yellow circle), 3 BE and T (orange points), and 23 T only (Red circle) as well as the low stringency reference (752) and polymorphic (218) insertions detected with L1-seq (green and orange respectively). Group contains 5 individuals.
C) Circos diagram mapping the distribution of 49 predicted high stringency insertions in tumor only (red circle), 537 low stringency reference (green circle), and 282 low-stringency polymorphic (orange circle) insertions. The final group contains 10 individuals. The somatic insertions (e.g. all insertions that are not reference and polymorphic insertions) are predictions and are therefore restricted by sequencing read counts of 100 reads or greater; however, the reference and polymorphic insertions have been previously published and are more common in the population and are therefore restricted only by a sequencing read count of 25 or greater and a map score of 0.5 or larger. For the Esophageal cancer group, our map scores were lower overall for the reference insertions; therefore, we restricted these insertions by a read count of 20 and a map score above 0.3.

Methods

L1-seq: DNA was isolated from the frozen tissue samples, from thinly sliced sections of tissue embedded in OTC freezing media with the DNeasy kit (Qiagen). Our samples were not micro-dissected to remove all normal tissue largely because half of our samples were either acquired as genomic DNA or previously frozen tissues. Equal amounts of genomic DNA from each individual were pooled by group. Hemi-specific PCR amplified the young, active L1 elements from the genome (249). Products between 200 and 500 sBowtie2, the alignments sorted based on presence or absence of L1 sequence, and the reference insertions and previously published polymorphic insertions were identified (249). Our bioinformatics analysis was essentially identical to previous analyses (249).

Stringency Analysis: For an insertion to be considered "high-stringency" in the library containing matched EAC and NE samples, we required at least a map score of 0.5 or greater, 50 total reads, and a window of 100 base-pairs or more spanning the junction of the 3' end of the L1 and the genomic DNA. Low-stringency insertions had below 50 reads, a map score of 0.5 or greater, and a window of less than 100 base pairs. These original parameters are similar to those previously used (249); however because we had more difficulty validating insertions in our other libraries we reevaluated the thresholds. For both of the remaining libraries: 1) the library containing the matched BE and NE samples and 2) the library containing matched EAC, BE, and NE samples we adjusted the thresholds by looking at the few validated insertions from the original high-stringency group. We noted the lowest unique read count, total read count, and window size among the previously validated insertions in each group, and used these numbers as our new parameters for high-stringency. We also required a higher map score for the redefined high-stringency insertions in both libraries. Consequently, a high-stringency insertion in the

library containing matched BE and NE samples required a map score of 0.8 or greater, 3 unique reads, 64 total reads, and a window of at least 107 base-pairs. High stringency insertions in the library containing matched EAC, BE, and NE samples required a least a map score of 0.8, 3 unique reads, 63 total reads, and a window of 140 base-pairs. For each insertion validated, the specific map score, read count, unique read count, and window size (bp) is noted (Table S3).

Random Insertion Selection: For the group of matched NE and EAC samples, insertions were randomly selected using a random number generator with parameters for both high and low stringency. A list of random numbers between 1 and the total number of predicted insertions (at varying levels of confidence) was then created and the rows which matched the numbers generated in the ".csv" file containing the predicted somatic insertions were selected for validation with PCR and sequencing. We made a histogram of the data to be sure the selection was even and random throughout the number range given and finally performed an empirical distribution analysis to evaluate our random selection process.

Immunohistochemistry: Immunohistochemistry was performed using the EnVision System-HRP (Dako; catalog K4006) according to the manufacturer's protocol. Primary antibody incubation was performed using the mouse monoclonal ORF1 (1.25 mg/ml) at a 1:3000 dilution for 40 min at room temperature. Secondary antibody incubation was performed per manufacturer protocol. For the skin biopsy, the sample was stained in an overnight protocol at a 1:1200 dilution with the monoclonal mouse ORF1 antibody. A second rabbit monoclonal ORF1 antibody was used to confirm initial results. This second antibody was used at a concentration of 1:2000 dilution with an overnight incubation at 4^{o} C and secondary antibody incubation as per manufacturer protocol. Orf1 monoclonal mouse antibody recognizes amino acids 35-44 while the

rabbit monoclonal antibody (JH74) detects the coiled-coil domain including amino acids 137-337 (209).

Results

BE patients without cancer

To estimate the pervasiveness of retrotransposition in BE, we studied 5 patients with BE who did not develop high-grade dysplasia or EAC for at least 15 years after their BE specimens were obtained. If L1 is active in patients without cancer this finding would suggest that the cellular environment in BE *per se* is permissive for retrotransposition. We obtained matched DNAs from white blood cells (WBC), NE, and BE and performed L1-seq to enrich DNA libraries for L1 insertions, then subsequently identify those insertions unique to the metaplasia (20). We classified these "somatic insertions" as those present only in a subset of cells and not inherited from a previous generation, e.g. insertions unique to BE but absent from matched NE and WBC DNA. Alternatively, we reasoned that somatic insertions could occur in a few normal squamous esophageal cells that became clonally amplified in BE. To confirm that an insertion was truly absent, we performed nested PCRs on all samples (Fig 2.2).

We confirmed a total of 20 insertions total in 4 of 5 patients evaluated by PCR and Sanger sequencing. Of the 20 confirmed insertions, 11 were amplified easily with a single PCR (conventional), without the need for a secondary PCR using nested primers (Fig 2.2A). We hypothesize that insertions which amplified with a conventional PCR were likely present in a large proportion of cells and therefore clonal. One insertion in particular was amplified easily with a conventional PCR in BE DNA; notably, this insertion was also observed in normal esophageal DNA only after nested PCR but remained undetectable in WBC DNA (Fig 2.3A-B). We speculated that this somatic insertion could have initially occurred in a single normal squamous cell exposed to high acid content during episodes of GERD, which then transdifferentiated into columnar epithelium, and clonally expanded as BE. This finding

suggests that L1 insertions occur in normal squamous esophageal cells at a low frequency and then become more easily detectable after they clonally expand in a disease such as BE or EAC, as previously suggested by Goodier for other tumor types(264).

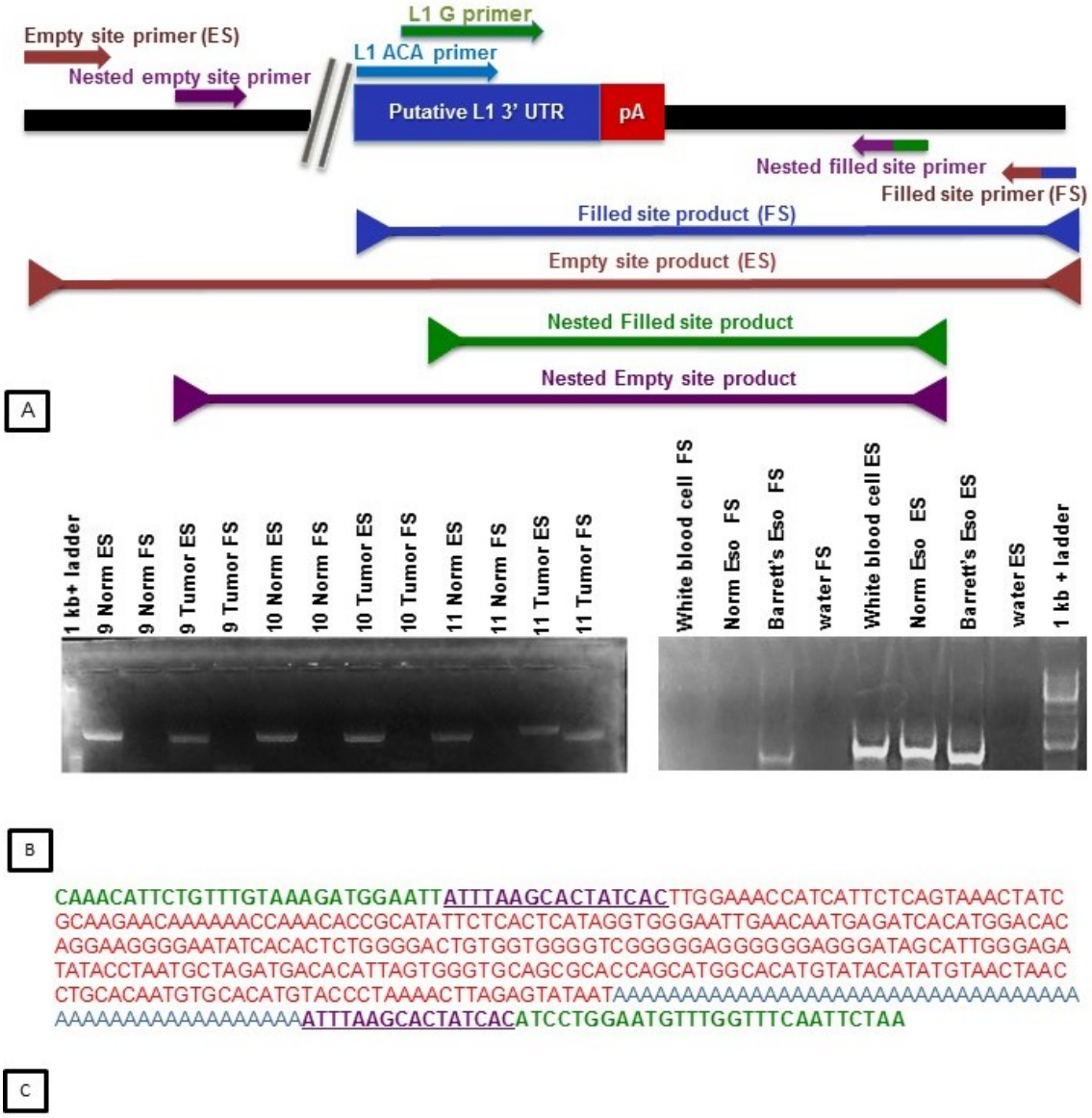


Figure 2.2: Somatic insertion validation process. (A) Diagram of the PCR validation scheme for putative insertions, the three prime end of the LINE-1 insertion is pictured adjacent to a poly A tail. The nested empty site and filled site primers are flanking the empty and filled site primers. In a nested PCR, the nested primers are used in the first of two reactions. Two uL of product from the first reaction (with ES and FS primers) is used as template in a second PCR with the nested primers to amplify difficult or rare products. (B) Two examples of validations for insertions present in only tumor and absent from normal DNA. On the left, inside the red box, a PCR result depicting both the empty site (ES) and filled site (FS) products for both the normal

and tumor DNA samples from patient 11. Only in the tumor is a filled site band present confirming the insertion is present in only the tumor DNA. In the image on the right side on Figure 2B, a PCR depicting another validation of a somatic insertion present in BE and absent from normal esophageal and white blood cell DNA. There is only a band present in the BE sample for the FS PCR; however, the ES PCR has bands for all three DNA samples as a positive control. (C) An insertion sequence with the unique genomic DNA(blue), target site duplications (purple), LINE-1 sequence (red), and the poly A tail sequence (orange).

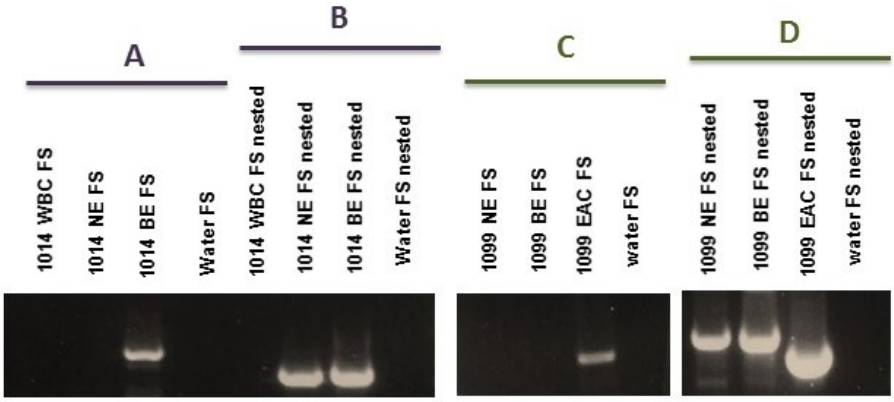


Figure 2.3: Gels showing clonal expansion of insertions originally present in NE. A) The first of two nested PCRs done on 1014 WBC, 1014 NE, 1014 BE DNA, and with a water control (no DNA), attempting to amplify the "Filled site" or the insertion in all three samples. There is only a band present in the BE DNA showing the insertion is likely clonal in BE. B) The second PCR or the "nested PCR" showing that with a nested PCR the insertion is present in 1014 NE too. C) The first of two nested PCRs done on 1099 NE, 1099 BE, 1099 T, and a water control (no DNA), attempting to amplify the "filled site" or the insertion in all three samples. A band is present for the PCR on the tumor DNA showing the insertion is likely clonal in the tumor. D) The second or "nested" PCR showing that with nested PCR the insertion is also present in NE and in BE DNA.

BE patients with cancer

After establishing that L1 is active in patients with benign BE, we evaluated individuals whose disease progressed to EAC. We hypothesized that individuals who develop EAC would have as many or more somatic insertion events due to increased genetic instability in frank cancer (6). We obtained samples from 5 patients with concomitant BE and EAC. Genomic DNA was isolated from NE, BE, and EAC tissues resected concurrently. After L1-seq we validated a number of these insertions in 2 of the 5 patients. We amplified and successfully Sanger sequenced 11 of 12 tested insertions which occurred in BE tissue alone, 27 of 36 in EAC alone, and 3 insertions that occurred in both BE and matched EAC. Due to the known polyclonal nature of BE, we had not expected all insertions detected in BE to be present in the matched EAC (262). We reasoned that typically, only one clonal population of cells should have evolved into the tumor and thus retained mutations acquired in the precursor lesion; the remaining clonal populations in the BE would not be expected to contain these same mutations (262).

The three insertions that were validated in multiple tissues provided a unique opportunity to look at the contribution of different clonal populations to the precursor lesion and the tumor. We observed three different stages at which a somatic insertion could occur. First, one of the insertions was detected without nested PCR in both BE and matched EAC; therefore, this insertion was likely part of a dominant BE clone which progressed to EAC. The second insertion was readily detected in EAC but required nested PCR to be detected in BE. Finally, a third insertion was amplified with conventional PCR in EAC, but was only evident in both NE and BE following nested PCR. This third insertion likely occurred in an NE cell, which evolved into BE, and subsequently clonally expanded in the EAC (Fig. 2.3 *C-D*). Altogether, these data

are further evidence that insertions occur at a low level in normal or metaplastic tissue and may later expand into a malignant clone. Similar observations, insertions which are easily amplified in the cancer tissue but only amplified in normal tissue following nested PCR reactions, have been observed by others in our lab in gastric cancer.

Similar to our previous group of non-progressive benign BE samples, only 2 of 5 individuals had somatic insertions in either BE, EAC, or both tissues. Although fewer patients had insertions in the matched BE-EAC group than in the group with BE or EAC alone, there were on average more somatic insertions validated in the patients with BE and EAC. In individuals with BE alone, we observed an average 5 insertions per person (20 insertions divided among 4 patients) while in patients with EAC there were 23.5 insertions on average per person (47 insertions among 2 patients). Because of the small sample size studied, it was impossible to determine whether this observed difference was statistically significant. Nevertheless, the wide range in the number of insertions per patient and the frequency of patients with insertions is in agreement with other observations (201–203,205,210).

EAC patients

To further investigate the activity of retrotransposons in EAC, we obtained samples from 10 additional patients with fresh-frozen matched NE and EAC tissue samples. Following L1-seq, we confirmed 49 of 72 randomly selected, high-stringency insertions (Methods) with PCR and Sanger sequencing. We then selected 20 low-stringency insertions (Methods) for validation and confirmed 6 additional somatic insertions. These confirmed insertions occurred in 4 of the 10 individuals' samples with great variation among individuals regarding the number of somatic insertions. Extrapolating from this large number of low-stringency predicted insertions by L1-seq

and our observed 30% validation rate in this group, we speculate that the number of potential L1 insertions per EAC is probably in the hundreds.

Previously, others have shown variability with respect to the number of confirmed somatic insertions per person as well as the proportion of individuals harboring somatic insertions (201–203,205,210). However, many studies have not thoroughly tested the potential clonality of the confirmed insertions. We tested 25 of the confirmed insertions in up to 6 tissue sections (20 in 6/6 sections and 2 in 2/2 sections) (Figure 2.4). We observed that 22 of these 25 insertions appeared in all sections tested, while the remaining 3 insertions were present in 5/6 sections tested (Figure 2.4). When insertions exist in multiple tissue sections, it suggests that they are likely clonal and may have occurred early during tumorigenesis, or even in the precursor lesion (BE). The concept of insertion clonality is important because it supports the conclusion that retrotransposition is active early during tumorigenesis.

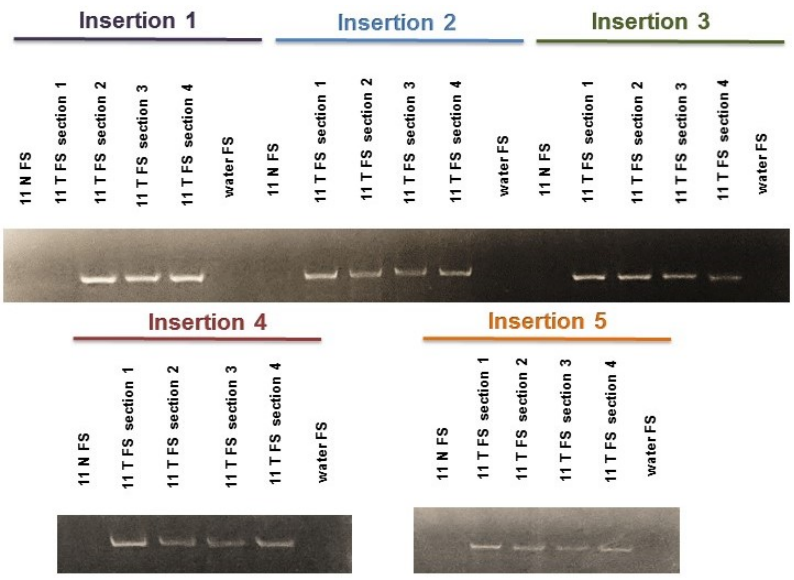


Figure 2.4: Representative gels illustrating the presence of specific insertions in multiple sections of tumor tissue. FS refers to the filled site PCR while ES refers to the empty site PCR as in figure 2.2. Two additional sections of each tumor were also tested and are not shown here.

Characterization of BE and EAC specific insertions

We established that retrotransposition is an active process in some BE and EAC patients by confirming 118 somatic insertions using PCR and Sanger sequencing. The confirmed insertions did not display an obvious bias for chromosomal location (Figure 2.1). In order to identify the precise insertion sites, we confirmed the 5' ends of 35 of the somatic insertions. For a subset of the insertions, we identified target-site duplications (TSDs) and endonuclease cleavage sites, both established hallmarks of retrotransposition (Table 2.1). However, out of 24 endonuclease cleavage sites identified, only 7 were similar (differed by two base pairs or less) to the canonical endonuclease site (111); the remaining 17 sites were more divergent from the canonical sequence. Furthermore, 11 insertions lacked TSDs and clear endonuclease cleavage sites indicating they were likely endonuclease independent insertions (254). Of the 11 insertions presumed endonuclease independent, 6 insertions had deletions at the site of integration (265).

Additional characteristics of these insertions, including mapping statistics of total read count, unique read count, and alignment windows, as well as genes nearby insertion sites are noted in Table 2.1. We observed variable lengths among the insertions for which we confirmed 5' ends, ranging from 111 to 1,579 nucleotides (without the poly-A tail) with 29 of 35 insertions, measuring under 500 nucleotides (Table 2.1). For the 21 of 35 insertions with TSDs, 12 were longer than 10 nucleotides (Table 2.1). One insertion contained a 3' truncated L1 element wherein 100 nucleotides of the 3' end of the L1 were deleted from the insertion site suggesting internal reverse transcriptase priming (45). Eleven of 35 insertions contained a 5' inversion, consistent with previous reports (201–203,205,210).

We did not confirm any insertions into exons in either BE or EAC; however, of the 118 confirmed somatic insertions, 48 insertions were into intronic regions of 50 genes. Twenty-three of these 48 insertions were into genes previously associated with cancer (Table S2). Our findings show a statistically significant enrichment of insertions into genes previously associated with cancer (p value < 1e-10 by Fisher's exact test). In order to accurately test for the observed enrichment, we accounted for the sizes of all the genes into which insertions occurred as well as the size of all cancer genes in the genome and the probability that an insertion would hit more of those genes by chance alone. After accounting for gene size, we still observed a significant enrichment of insertions into genes previously associated with cancer (p value < 1e-10).

Insertion Number	Chr	Position	Group	Disease	Sample	Sections present	In Gene	unique reads	total reads	width	unique reads T	total reads T	width T	map score
1	1	116047929	1	BE	1014B		-----	16	1987	496				1
2	1	19007407	1	BE	1014B		PAX7	7	605	192				1
3	4	180952222	1	BE	1089B		-----	11	2849	108				1
4	8	26755602	1	BE	1089B		-----	4	64	149				1
5	5	3808832	1	BE	1089B		-----	7	142	266				1
6	11	15280187	1	BE	1014B		-----	6	100	277				1
7	8	64992514	1	BE	1014B		LOC102724623	8	905	413				1
8	18	10353659	1	BE	1014B		-----	3	71	107				1
9	3	113012552	1	BE	1014B		C3orf17,WDR52	3	97	118				1
10	10	124528777	1	BE	1014B		DMBT1P1	7	86	359				1
11	8	14924043	1	BE	1014N* and 1014B		SGCZ	6	85	419				1
12	8	20956979	1	BE	1014B		-----	4	83	257				1
13	2	19437601	1	BE	1014B		-----	7	65	442				1
14	2	19241638	1	BE	1014B		-----	5	131	169				1
15	5	117055085	1	BE	1014B		-----	4	99	164				1
16	8	5992496	1	BE	1014B		-----	6	69	293				1
17	8	131156942	1	BE	1094B		ASAP1	10	241	197				1
18	18	11258010	1	BE	1014 B		-----	13	254	233				1
19	11	40419074	1	BE	1094 B		LRRC4C	12	185	267				1
20	8	40629224	1	BE	809 B		ZMAT4	7	214	304				1
21	1	204737046	2	EAC/BE	1099B		-----	7	907	230				1
22	2	175708654	2	EAC/BE	1099B		CHN1	18	615	301				1
23	4	112345417	2	EAC/BE	1099B		-----	10	285	224				0.833
24	4	137020965	2	EAC/BE	1099B		-----	7	651	187				1
25	9	19758216	2	EAC/BE	1099B		SLC24A2	6	195	188				1

Insertion Number	Chr	Position	Group	Disease	Sample	Sections present	In Gene	unique reads	total reads	width	unique reads T	total reads T	width T	map score
26	9	83216370	2	EAC/BE	1498B		-----	4	567	167				1
27	16	5717356	2	EAC/BE	1099B		-----	15	208	226				1
28	18	63294543	2	EAC/BE	1099B		-----	10	454	339				1
29	X	65605347	2	EAC/BE	1099T		-----	27	419	467				1
30	3	166726115	2	EAC/BE	1099T		-----	12	4776	364				1
31	8	69568706	2	EAC/BE	1099T		C8ORF34	36	8908	556				1
32	X	50596639	2	EAC/BE	1099T		-----	12	185	547				1
33	12	94347882	2	EAC/BE	1099T		-----	40	425	312				1
34	1	37600548	2	EAC/BE	1099T		-----	42	631	547				1
35	1	232267247	2	EAC/BE	1099B		-----	11	1709	279				1
36	3	47159145	2	EAC/BE	1099B		SETD2	7	472	401				1
37	3	77236703	2	EAC/BE	1099B		ROBO2	12	298	327				1
38	6	97954428	2	EAC/BE	1099B and 1099T	2/2	LOC101927314	10	557	334	1	1	90	0.893
39	14	88959897	2	EAC/BE	1099B		PTPN21	11	234	319				1
40	13	59989835	2	EAC/BE	1099B		-----	7	204	165				1
41	14	44376719	2	EAC/BE	1099B		-----	8	296	143				1
42	8	140448417	2	EAC/BE	1099B		-----	11	31005	211				1
43	2	41842730	2	EAC/BE	1099B* and 1099T	2/2	-----	13	2196	220	4179	21	220	1
44	3	147275713	2	EAC/BE	1099B		-----	19	1436	250				1
45	13	71325493	2	EAC/BE	1498B		-----	11	223	264				1
46	1	170042996	2	EAC/BE	1498B		KIFAP3	13	68	317				1
47	3	70674092	2	EAC/BE	1498B		-----	13	395	342				1
48	5	137341463	2	EAC/BE	1498B		FAM13B	10	882	186				1
49	1	114605790	2	EAC/BE	1099T		-----	3	63	140				1
50	2	2622020	2	EAC/BE	1099B		----	6	316	147				1

Insertion Number	Chr	Position	Group	Disease	Sample	Sections present	In Gene	unique reads	total reads	width	unique reads T	total reads T	width T	map score
51	8	110325236	2	EAC/BE	1099B		NUDCD1	12	796	387				1
52	11	24820370	2	EAC/BE	1099B		LUZP2	7	267	167				1
53	5	161834035	2	EAC/BE	1099T		-----	15	208	584				1
54	6	133354504	2	EAC/BE	1099T		-----	20	266	552				1
55	8	35312419	2	EAC/BE	1099T		-----	22	2716	540				1
56	1	63399166	2	EAC/BE	1099N*, 1099B*, 1099T		-----	28	1079	516				1
57	7	152811588	2	EAC/BE	1099T		-----	17	381	516				1
58	2	69239926	2	EAC/BE	1099T		ANTRX1	24	489	510				1
59	14	30286674	2	EAC/BE	1099T		PRKD1	18	581	498				1
60	6	28522512	2	EAC/BE	1099T		----	16	1142	471				1
61	5	7471214	2	EAC/BE	1099T		ADCY2	24	310	450				1
62	Y	22673889	2	EAC/BE	1099T		TTTY10	15	2319	446				1
63	14	80879070	2	EAC/BE	1099T		DIOS-AS1	27	174	415				1
64	X	120351391	2	EAC/BE	1099T		-----	15	428	389				1
65	6	39701567	2	EAC/BE	1099T		-----	10	185	381				1
66	5	33689981	2	EAC/BE	1099T		-----	30	9050	375				1
67	7	90764574	2	EAC/BE	1099T		CDK14	52	1786	367				1
68	7	14722528	2	EAC/BE	1099T		-----	13	713	323				1
69	12	94347882	2	EAC/BE	1099T		----	40	425	312				1
70	16	78512648	3	EAC	11T	5/6	WWOX	9	65	445				1
71	2	230359765	3	EAC	11T	6/6	DNER	8	113	265				0.785
72	4	187797210	3	EAC	11T		----	3	74	169				1
73	2	56419065	3	EAC	11T	6/6	CCDC85A	12	1194	265				1
74	2	197718019	3	EAC	11T		PGAP1	20	408	216				1
75	15	42034587	3	EAC	11T		MGA	3	52	338				1

Insertion Number	Chr	Position	Group	Disease	Sample	Sections present	In Gene	unique reads	total reads	width	unique reads T	total reads T	width T	map score
76	2	44848531	3	EAC	11T	6/6	CAMKMT	17	660	272				1
77	13	55308060	3	EAC	11T	6/6	-----	5	88	190				1
78	13	71024272	3	EAC	17T	6/6	-----	2	102	328				1
79	11	7305164	3	EAC	11T	6/6	SYT9	5	272	108				1
80	2	17905038	3	EAC	11T	6/6	SMC6	18	220	434				1
81	17	55278400	3	EAC	11T	5/6	-----	13	445	238				1
82	2	131825239	3	EAC	11T	6/6	FAM168B	8	65	115				1
83	2	145568629	3	EAC	11T	6/6	TEX41	14	254	337				1
84	8	122208109	3	EAC	11T		-----	6	133	487				1
85	2	135596396	3	EAC	11T	6/6	ACMSD	6	166	237				1
86	4	85357755	3	EAC	12T		-----	8	635	389				1
87	6	77062721	3	EAC	11T		-----	6	104	176				1
88	6	46356921	3	EAC	11T	6/6	RCAN2	5	82	194				1
89	11	114846922	3	EAC	11T		-----	12	113	259				1
90	12	10022376	3	EAC	11T		CLEC2B	5	70	128				1
91	15	48171526	3	EAC	11T		-----	7	183	207				1
92	8	133385057	3	EAC	11T		KCNQ3	15	1178	329				1
93	8	77093391	3	EAC	10T	6/6	-----	6	299	312				1
94	16	56494481	3	EAC	11T		OGFOD1	6	421	342				1
95	5	125531877	3	EAC	11T	5/6	-----	5	152	144				1
96	8	55529895	3	EAC	10T	6/6	RP1	10	978	181				1
97	6	98636002	3	EAC	10T		-----	9	341	234				1
98	12	99893559	3	EAC	11T	6/6	ANKS1B	9	147	413				1
99	11	94929580	3	EAC	11T		SESN3	12	465	227				1
100	4	27638735	3	EAC	10T		-----	4	128	452				1

Insertion Number	Chr	Position	Group	Disease	Sample	Sections present	In Gene	unique reads	total reads	width	unique reads T	total reads T	width T	map score
101	1	164209317	3	EAC	11T	6/6	-----	11	1373	105				1
102	10	133295664	3	EAC	11T	6/6	-----	8	135	259				1
103	X	146043349	3	EAC	11T		-----	3	61	117				1
104	6	45309543	3	EAC	11T	6/6	RUNX2/SUPT3H	13	106	333				1
105	3	102069771	3	EAC	11T		-----	12	699	476				1
106	X	110754564	3	EAC	11T	6/6	LINC00890	4	65	273				1
107	1	216454770	3	EAC	11T		USH2A	4	86	377				1
108	5	104118195	3	EAC	11T		-----	35	2812	462				1
109	7	142025178	3	EAC	11T		TCRBV5S1A1099T	22	1728	182				1
110	5	125531877	3	EAC	11T		-----	5	152	144				1
111	3	191665653	3	EAC	11T		-----	4	80	201				0.833
112	5	135931565	3	EAC	11T		-----	2	49	93				1
113	16	5933497	3	EAC	10T	6/6	-----	3	42	97				1
114	12	45518842	3	EAC	11T		-----	1	30	90				1
115	10	68442928	3	EAC	11T		CTNNA3	8	383	238				1
116	5	4556686	3	EAC	11T		-----	2	27	91				1
117	5	87131402	3	EAC	10T		-----	3	31	96				1
118	X	125740689	3	EAC	10T	6/6	-----	2	31	91				1

Table 2.1: Somatic Insertion Characterization. All validated insertion locations noted in the table, along with TSDs, endonuclease cleavage sites, poly A tail length, and introns of genes into which insertions occurred. The asterisks denote samples in which insertions were detected only with a nested PCR.

Gene Name	Cancer type(s)	Reference(s)	NGC 4.0	Entrez Gene
LUZP2	chronic lymphocitic leukemia	Quesada V, Conde L, Villamor N, et al. (2011) Exome sequencing identifies recurrent mutations of the splicing factor *SF3B1* gene in chronic lymphocytic leukemia. *Nature Genetics* 44(1): 47-52.	Y	338645
PRKD1	colorectal	Wood L, Parsons D.W., Jones, S, et al. (2007) The Genomic Landscapes of Human Breast and Colorectal Cancers. *Science* 318(5853):1108-1113.	Y	5587
ADCY2	lung	Lee W, Jiang Z, Liu J, et al. (2010) The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* 465(7297):473-477.	Y	108
PAX7	sarcoma	Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster, Rahman N, and Stratton MR. (2004) A census of human cancer genes. *Nature Rev Cancer* 4(3):177-183.	Y	5801
SETD2	kidney, breast	Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster, Rahman N, and Stratton MR. (2004) A census of human cancer genes. Nature Rev Cancer 4(3):177-183. Stephens PJ, Tarpey PS, DaviesH, et al. (2012) The landscape of cancer genes and mutational processes in breast cancer. Nature 486(7403):400-404. Guo G, Gui Y, Gao S, et al. (2011). Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. Nature Genetics 44(1):17-19. Dalgliesh GL, Furge K, Greenman C, et al. (2010) Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature 463(7279):360-363.	Y	29072

Gene Name	Cancer type(s)	Reference(s)	NGC 4.0	Entrez Gene
ROBO2	cholangiocarcinoma	Onk CK, Subimerb C, Pairojkul C, et al. (2012). Exome Sequencing of liver fluke associated cholangiocarcinoma. Nature Genetics 44(6):690-693.	Y	6092
DNER	breast, lung, ovarian, pancreas, prostate	Berger MF, Lawrence MS, Demichelis F, et al. (2011) The genomic complexity of primary human prostate cancer. Nature 470(7333):214-220. Kan Z, Jaiswal BS, Stinson, J, et al. (2010) Diverse somatic mutation patterns and pathway alteration sin human cancers. Nature 466(7308):869-873.	Y	92737
CCDC85A	breast	Ding L, Ellis MJ, Larson DE, et al. (2010) Genome remodeling in a basal-like breast cancer metastasis and xenograft. *Nature* 464(7291):999-1005.	Y	114800
MGA	chronic lymphocitic leukemia	Puente XS, Pinyol M, Quesada V, et al. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature* 475(7354):101-105.	Y	23269
KCNQ3	prostate	Berger MF, Lawrence MS, Demichelis F, et al. (2011) The genomic complexity of primary human prostate cancer. *Nature* 470(7333):214-220.	Y	3786
RP1	breast, melanoma	Shah SP, Morin RD, Khattra J, et al. (2009) Mutational evolution in a lobular breast tumour profiled at a single nucleotide resolution. Nature 461(7265):809-813. Turajlic S, Furney SJ, LAmbros MD, et al. (2012) Whole genome sequencing of matched primary and metastatic acral melanomas. Genome Research 22(2):196-207.	Y	6101

Gene Name	Cancer type(s)	Reference(s)	NGC 4.0	Entrez Gene
USH2A	breast, head and neck squamous cell carcinoma, pancreas	Koboldt DC, Fulton RS, McLellan MD, et al. (2012) Comprehensive molecular portraits of human breast tumours. Nature 490(7418):61-70. Stansky N, Egloff AM, Tward AD, et al. (2011) The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. Science 333(6046):1157-1160. Wang L, Tsutsumi S, Kawaguchi T, et al. (2012) Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome Research 22(2):208-219. Koboldt DC, Fulton RS, McLellan MD, et al. (2012) Comprehensive molecular portraits of human breast tumours. Nature 490(7418):61-70. Stansky N, Egloff AM, Tward AD, et al. (2011) The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. Science 333(6046):1157-1160. Wang L, Tsutsumi S, Kawaguchi T, et al. (2012) Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome Research 22(2):208-219.	Y	7399
CTNNA3	glioblastoma	Clark MJ, Homer N, O'Conor BD, et al. (2010) U87MG Decoded: The Genomic Sequence of a Cytogenetically Aberrant Human Cancer Cell Line. *PLOS Genetics* 6(1):e1000832.	Y	29119
CHN1	sarcoma	Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster, Rahman N, and Stratton MR. (2004) A census of human cancer genes. Nature Rev Cancer 4(3):177-183.	Y	1123
CLEC2B/ AICL	cholangiocarcinoma	Wang AG, Yoon SY, Oh JH, et al. (2006) Identification of intrahepatic cholangiocarcinoma related genes by comparison with normal liver tissues using expressed sequence tags. IBiochem Biophys Res Commun 345(3):1022-1032. 52. Akatsuka A, Ito M, Yamauchi C, Ochial A, Yamamoto K, Matsumoto N. (2010) Tumor cells of non-hematopoietic and hematopoietic origins express activation-induced C-type lectin, the ligand for killer cell lectin-like receptor F1. Int Immunol 22(9):783-790. 53.	N	9976

Gene Name	Cancer type(s)	Reference(s)	NGC 4.0	Entrez Gene
ANKS1B	leukemia	Ghersi E, Vito P, Lopez P, Abdallah M, and D'Adamio L. (2004) The intracellular localization of amyloid beta protein precursor (AbetaPP) intracellular domain associated protein-1 (AIDA-1) is regulated by AbetaPP and alternative splicing. J Alzheimers Disease 6(1):67-68. Casagrande, G, te KRonnie G, and Basso G. The effects of siRNA-mediated inhibition of E2A-PBX1 on EB-1 and Wnt16b expression in the 697 pre-B leukemia cell line. Haematologica 91(6):765-771.	N	56899
RUNX2	osteosarcoma	San Martin IA, Varela N, Gaete M, et al. (2009) Impaired cell cycle regulation of the osteoblast-related heterodimeris transcription factor Runx2-Cbfbeta in osteosarcoma cells. *J Cell Physiol* 221(3):560-571. 56. van der Deen M, Akech J, Lapointe D, et al. (2012) Genomic promoter occupancy or runt-related transcription factor RUNX2 in Osteosarcoma cells identifies genes involved in cell adhesion and motility. J Biol Chem 287(7):4503-4517.	N	860
ASAP1	colorectal cancer	Muller T, Stein U, Poletti A, et al. (2010) ASAP1 promotes tumor cell motility and invasiveness, stimulates metastasis formation in vivo, and correlates with poor survival in colorectal cancer patients. *Oncogene* 29(16):2393-2403.	N	50807
WWOX	carcinoma squamous cell and epithelial tumors	Lai, F, Cheng CL, Chen ST, Wu CH, Hsu LJ, Lee JY, Chao SC, Sheen MC, Shen CL, Chang NS, and Sheu HM. (2005) WOX1 is essential for UVB irradiation-induced apoptosis and down-regulated via translational blockade in UVB-induced cutaneous squamous cell carcinoma in vivo. Clin Cancer Res 11(16):5769-5777. Ishii H, Mimori K, Vecchione A, Sutheesophon K, Fujiwara T, Mori M, and Furukawa Y. (2004) Effect of exogenous E2F-1 on the expression of common chromosome fragile site genes, FHIT and WWOX. Biochem Biophys Res Commun 316(4):1088-1093.	N	51741

Gene Name	Cancer type(s)	Reference(s)	NGC 4.0	Entrez Gene
SMC6	osteosarcoma	Potts PR and Yu H. (2007) The SMC5/6 complex maintains telomere length in ALT cancer cells through SUMOylation of telomere-binding proteins. Nat Struct Mol Biol 14(7):581-590.	N	79677
LRRC4C	extragonadal seminoma and gastric cancer	Vauhkonen H, Vauhkonen M, Sipponen P, and Knuutila S. (2007) Oligonucleotide array comparative genomic hybridization refines the structure of 8p23.1, 17q12 and 20q13.2 amplifications in gastric carcinomas. Cytogenet Genome Res 119 (1-2):39-45. University of Copenhagen Diseases database linked to extagonadal seminoma	N	57689
PTPN21	colorectal cancer	Korff S, Woerner SM, Yuan YP, Bork P, von Knebel Doeberitz M, and Gebert J. (2008) Frameshift mutations in coding repeats of protein tyrosine phosphatase genes in colorectal tumors with microsatellite instability. BMC Cancer 329 doi: 10.1186/1471-2407-8-329.	N	11099
DMBT1P1	breast cancer	Blackburn AC, Hill LZ, Roberts AL, et al. (2007) Genetic mapping in mice identifies DMBT1 as a candidate modifier of mammary tumors and breast cancer risk. AM J Pathol 170 (6):2030-2041.	N	375940

Table 2.2: Confirmed Somatic Insertion into Genes Associated with Cancer. Cancer genes into which confirmed insertions occurred, relevant references for cancer association, and entrez gene ID. NCG 4.0 column indicates whether or not the cancer gene is included in the network of cancer genes available at ngc.kcl.ac.uk/.

L1 Expression in NE and EAC

Theoretically, retrotransposition is dependent on L1 protein expression for its activity; therefore, the genetic evidence of somatic insertions strongly suggests L1 proteins are expressed in precancerous lesions as well as cancer. Expression of L1 protein has been observed in many cancer types previously but was only rarely detected in histologically normal tissue adjacent to the cancer (209,266,267). Furthermore, the evidence of somatic insertions in normal esophagus suggests there must be at least transient or a low-level of L1 protein expression in the tissue. One of the two proteins encoded by L1, open-reading frame 1p (ORF1p), has been observed in many cancer types and has occasionally been observed in normal tissue adjacent to cancer (209,266,267). To evaluate ORF1p expression in the patients harboring somatic L1 insertions, we obtained formalin-fixed paraffin-embedded (FFPE) tissues from 8 of the aforementioned EAC patients.

We observed ORF1p expression in all 8 of these tumor samples by immunohistochemistry (IHC) (Table S3). The level of ORF1p expression varied among individuals as well as within individual tissue sections where cancerous glands were developing (Figure 2.5A-F). All of the samples with a confirmed somatic insertion showed ORF1p expression. There was no correlation between protein expression and the number of confirmed somatic insertions per individual (Table S3). Interestingly, we detected low level ORF1p expression in all 4 of the available matched normal tissues in both the stratified squamous epithelium and the smooth muscle (Figure 2.6A-F). Expression was absent from the progenitor stem cells of the stratified squamous epithelium and seemed to increase with cellular maturation as the cells increased their cytoplasm and radiated away from their progenitors. The expression was absent from the submucosa of the tissue. Expression was observed with two separate

monoclonal antibodies that detect different epitopes of the protein (209). Although it was surprising to observe ORF1p expression in normal esophagus, it does support our finding of somatic insertions in normal esophagus that later expanding in subsequent metaplasia and/or cancer.

To investigate whether the expression present in the normal esophagus was limited to patients who had concomitant cancer, we obtained one normal esophagus sample from a biopsy conducted on a patient with gastric ulcers. We also obtained a normal skin biopsy to evaluate the squamous epithelium expression of ORF1p in an epithelial tissue. In both the normal biopsies, dim ORF1p immunoreactivity was evident in the squamous epithelium of the tissue (Figure S1A-B). LINE-1 expression in normal epithelial tissues, albeit at low levels, may allow for retrotransposition events. Perhaps a subset of somatic retrotransposition events reported in epithelial cancers actually occur prior to transformation (201–203,205,210). At the same time, the higher levels of LINE-1 expression we see in these cancers may selectively promote somatic insertion events in malignant cells.

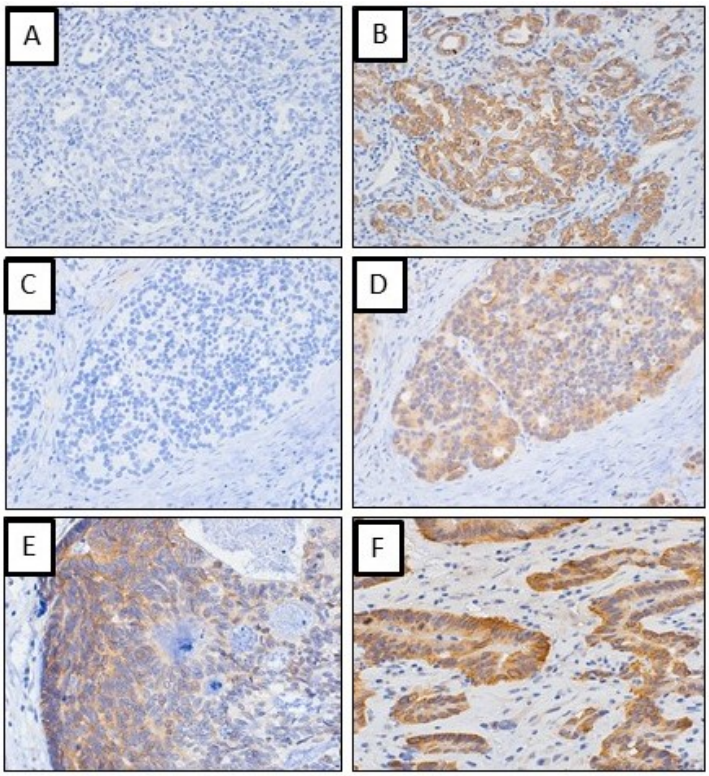


Figure 2.5: Representative photomicrographs depicting LINE-1 ORF1p immuno-labeling in esophageal carcinomas. A) Virtually no immuno-labeling identified with the primary LINE-1 ORF1p antibody was not used in the IHC procedure (*i.e.* no antibody control) Final magnification x100 B) Same case as A, when incubated with LINE-1 ORF1p antibody, indicating the cancer is strongly reactive for ORF1p antigen. Final magnification x100 C) Virtually no immuno-labeling identified in the no antibody control. Final magnification x100. D) Same case as C, when incubated with LINE-1 ORF1p antibody, indicating the cancer is reactive for ORF1p antigen. Final magnification x100. E)-F) Two additional EAC cases which are reactive for ORF1p antigen. Final magnification x160.

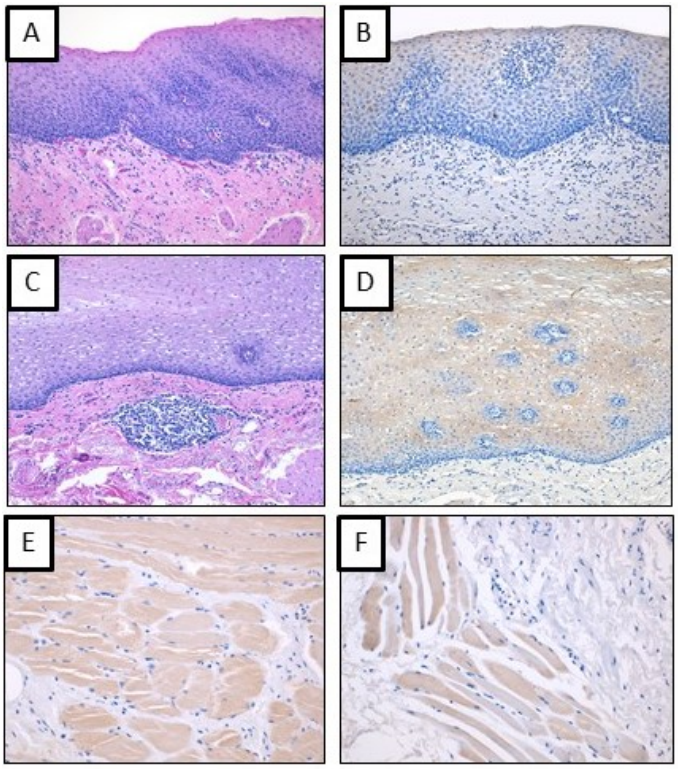


Figure 2.6: Representative photomicrographs depicting LINE-1 ORF1p expression in normal esophageal tissue. A) and C) Normal esophageal tissue from two distinct individuals stained with H&E final magnification x100. B) Same case as A) when incubated with LINE-1 ORF1p antibody, indicating the normal esophageal tissue is reactive for ORF1p antigen. Final magnification x100. D) same case as C) when incubated with LINE-1 ORF1p antibody, indicated the normal esophageal tissue is reactive for ORF1p antigen. Final magnification x100. E) and F) Normal esophageal tissue from two distinct individuals showing the smooth muscle is reactive for ORF1p antigen. Final magnification x160.

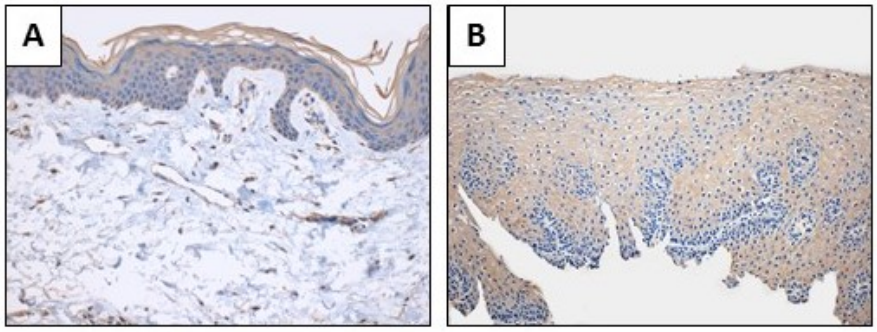


Figure 2.7: ORF1p expression in normal esophagus and normal skin samples. A) Normal skin biopsy showing ORF1p expression localized in the stratified squamous epithelium. Final magnification x160. B) Normal esophagus biopsy from patient with stomach ulcers showing ORF1p expression localized in the stratified squamous epithelium. Final magnification x100.

Patient	Disease	Level of Expression	Insertions Validated
9	EAC	++	0
10	EAC	+	7
11	EAC	++	40
12	EAC	+	1
14	EAC	+	0
16	EAC	++	0
17	EAC	+	1
18	EAC	+	0

Table 2.3: Correlation between ORF1p expression and somatic insertions in EAC. Table displaying relationships between ORF1p expression and insertion occurrence in all subjects with both L1-seq and IHC data.

Discussion

Improved understanding of carcinogenesis should lead to earlier diagnosis and more effective treatment, but this advance requires the study of precursor lesions. In many ways, BE is ideal for studying clonal expansion in precursor lesions even when disease progression does not occur. BE is accessible and present in a sizable proportion of the Western population, even though it only progresses to EAC in a small subset of patients (259). Cellular processes that are dysregulated during the transdifferentiation from stratified squamous epithelium to metaplastic columnar epithelium may provide a fertile environment for dysregulation of L1.

We found that retrotransposition is active in a subset of individuals with BE and EAC; however, this process does not occur in all patients and is active in patients with long-standing benign disease. Therefore, L1 activity alone is not a reliable predictor of disease progression in BE. BE appears to provide a permissive environment for L1 retrotransposition, which in turn increases the mutational burden and potentially contributes to disease progression. As evidence of this permissive environment, we demonstrated that L1 elements were active in 6 of the 10 BE tissues evaluated by confirming 46 new somatic insertions. Furthermore, we validated 75 insertions in 6 of 15 EAC samples. Where somatic insertions occurred, there was a variable frequency of events, ranging from 1 to 44 insertions, among different individuals. In contrast to our previous colon cancer study (201), we did not observe a linear correlation between the number of insertions and any other characteristic, including age or L1 protein expression.

Many of the insertions validated in BE and EAC had characteristics which differ from typical germline somatic insertions (51). First, we observed 8 of 35 (23%) insertions with integration-site deletions, much greater than the 10% seen in the germline (111). Our failure to detect more than 30% of the 5' ends may, in a number of the cases, be due to even larger

integration-site deletions (265). Secondly, we found 11 insertions which appeared to be endonuclease independent, a much larger number than that observed in germline insertions (111,249). These insertions were presumed to be endonuclease independent due to their lack of both target site duplications and clear endonuclease cleavage sites, both hallmarks of the canonical process of retrotransposition. Thirdly, the majority of the insertions for which we identified the 5' end were highly truncated with 29 of 35 below 500 nucleotides in length. Also, we did not detect any 3' transductions among our confirmed insertions in contrast to the findings of others (202), but L1-seq detects only a small fraction of these events.

Somatic L1 insertions are seldom observed in normal tissues (221) with the most notable exception being those observed in the hippocampus (220,268). Between our three groups of samples analyzed with L1-seq, we attempted to validate 9 high-stringency insertions predicted in normal esophagus only. Even with nested PCR we were unable to confirm any of these normal-only insertions, a result that we have seen previously for normal-specific insertions (203). Interestingly, we validated two insertions in normal tissue that were also present in BE and EAC. This finding suggests that at least some insertions may occur in normal squamous epithelium cells and are then selected for in the ensuing pathological state. We speculate that indeed many of the BE and EAC insertions occur initially in only one or a small number of normal esophageal cells. Clonal expansion in diseases such as BE and EAC may make it easier to detect the low level of L1 activity in a subset of normal cells. Future studies utilizing single-cell sequencing may allow us to better determine the activity of L1 in normal tissues and whether insertions in BE or in tumor are truly clonal.

Somatic retrotransposition occurs at a detectable rate in squamous cell lung, head and neck, colorectal, endometrial, hepatocellular, breast, prostate, bone, and various other types of

cancer (201–203,205,210). We now demonstrate that this process occurs in premalignant BE and EAC. Although retrotransposition does not occur in all BE and EAC patients and recurrent insertions were not found, L1 may still participate in carcinogenesis. It appears that although epithelial cancers are permissive for retrotransposition, there may be other factors mediating this process that allow it to occur in certain individuals more than in others. Identifying the factors underlying the activation of retrotransposition, as well as the contributions it makes to carcinogenesis, will be essential to improve our understanding of genomic instability generated by L1 and the role of retrotransposition in epithelial tumor development.

CHAPTER 3:

LINE-1 Expression and Retrotransposition in Normal Esophagus and Esophageal Squamous Cell Carcinoma

Abstract

Squamous cell carcinoma of the esophagus (SCC) is the most common form of esophageal cancer in the world and is frequently diagnosed at an advanced stage when successful treatment is challenging. Understanding the mutational profile of cancer is necessary to develop robust and successful treatments. Because many groups, including our own, observed somatic retrotransposition in tumors of the digestive system, we focused on LINE-1 (L1) mobilization as a source of instability in this cancer. We hypothesized that retrotransposition is ongoing in SCC patients. The expression of L1 proteins is necessary for active retrotransposition to occur; therefore, we evaluated the expression of open reading frame 1 protein (ORF1p), a protein encoded by L1. Using immunohistochemistry we detected ORF1p expression in all four of the nine available SCC cases. After L1-seq, we confirmed 74 somatic insertions in the tumors of eight of nine individuals evaluated. Interestingly, we found 12 insertions that appeared to be sub-clonal in the adjacent normal esophagus while likely clonal in the tumor using both conventional and nested PCR. Overall, our results indicate that L1 retrotransposition is active in squamous cell carcinoma of the esophagus and that early events occurring in histologically normal esophagus may frequently expand clonally in the resulting tumor.

Introduction

Squamous cell carcinoma of the esophagus (SCC) is the most common esophageal cancer in the world and its incidence differs across various geographic areas; its incidence is low in North America but it commonly occurs in parts of Eastern Asia (269,270). SCC develops from the cells comprising the squamous esophageal mucosa and its main risk factors include combined alcohol and tobacco use, consumption of scalding beverages, and a diet low in fresh fruits and vegetables (269,271). Esophageal squamous cell carcinoma is the major histologic type of esophageal cancer in East Asian countries and is an aggressive tumor (272). The cancer is especially common in rural, mountainous areas with little access to resources and minimal dietary diversity such as Northern Iran, central China, parts of South-East Africa, and South America (271,273–275). Unfortunately, by the time SCC is diagnosed, greater than half of the patients have inoperable tumors or obvious metastases (272). Even if the tumor is removable, the prognosis for most patients is still very poor; therefore, better methods for early detection and treatment are necessary (276).

Recently many groups, including our own, have evaluated a mutation generating mechanism known as retrotransposition and its role in epithelial cell cancer development (201–206,208,210). During retrotransposition, a sequence of DNA mobilizes via an RNA mediated "copy and paste" mechanism. L1 elements are the only autonomous retrotransposons in the human genome. These elements mobilize themselves by promoting their own transcription followed by the translation of the two open reading frames coding for proteins necessary for the element's reintegration into the genome. The movement of all other retro-elements in the human genome (Alu, SVA, and processed pseudogenes) is dependent on the activity of L1 elements. Because retrotransposons are potentially mutagenic when inserting into new sites in the genome,

host cells inhibit their activity by suppressing transcription and translation (60,61,63,74,75,78,79,85,277,278). L1 expression levels are inversely correlated with methylation of the promoter in the 5' UTR of the element and numerous epigenetic modifiers contribute to establishing and maintaining the methylation status of L1 elements in the genome (68,71,279,280). A known feature of SCC is hypomethylation of L1 elements throughout the genome; furthermore, the less methylation present, the worse the prognosis for the patient (281). Additionally, we have observed L1 protein expression of ORF1, open reading frame 1, in normal squamous epithelium of the esophagus indicating the L1 is potentially active in the relevant normal tissue (206).

We hypothesized that in some individuals, L1 is active in a subset of cells in the normal squamous epithelium and its subsequent insertions may contribute to the process of esophageal squamous cell tumor development. When one or more of the host control mechanisms fails, L1 is capable of mobilizing and increasing the mutational burden of the genome. In cancer, genomic instability and hypomethylation may create a hospitable environment for L1 expression and mobilization. Therefore, we expected that in SCC, a cancer known to be hypo-methylated at L1 promotors, L1 elements would express proteins more robustly and increase activity. In a previous study, we demonstrated that a sub-clonal insertion occurred in the histologically normal tissue of the esophagus and was expanded in the subsequent metaplastic condition of a patient (206). We also observed a sub-clonal insertion which occurred in histologically normal tissue, was maintained in the population of metaplastic cells, and then expanded in the subsequent esophageal adenocarcinoma (206).

To test our hypothesis, we evaluated L1 mobilization in individuals with esophageal squamous cell carcinoma using L1-seq, a high-throughput L1 targeted next-generation

sequencing method (249). We observed L1 activity in SCC and in NE and demonstrated that sub-clonal insertions occur in normal squamous epithelium at a higher rate than previously observed in other cancer types (204,206).

L1-seq detected reference, non-reference, and somatic insertions in SCC patients							
Disease	**Group**	**Number of Individuals**	**SCC patients**	**Known reference**	**Known non-reference**	**Reads required**	**Map score**
SCC/EAC	1	5	4	1005	350	10	0.3
SCC	2	5	5	640	253	10	0.3
Disease	**Group**	**Number of Individuals**	**SCC patients**	**Patients with insertions**	**Tumor-only**	**NE sub-clonal/SCC clonal**	**Normal only**
SCC/EAC	1	5	4	4	17	0	0
SCC	2	5	5	4	48	12	0

Table 3.1: This table shows the overall number of validated somatic insertions in each group of patients. It also contains the metrics for L1-seq, namely, the number of reference and non-reference insertions detected with next-generation sequencing for each library which give an estimate of the coverage of the technique.

Materials and Methods

L1-seq. Using the DNeasy kit from Qiagen, we isolated DNA from thinly shaved sections of fresh-frozen tissue embedded in OTC freezing media. Five pairs of our patient samples were micro-dissected with the assistance of pathologist Robert A. Anders (including samples: 20, 21, 22, 23, and 24 N and T). The microdissection removed all normal tissue from the tumor, and all tumor tissue from the normal, in addition to removing areas where necrosis was evident. Equal amounts of genomic DNA from each individual were pooled by group (either normal or tumor) in the same manner previously used (206). L1-seq uses a hemi-specific PCR and eight degenerate primers to selectively amplify human-specific active L1 elements in the human genome (249). Following the PCR phase of L1-seq, we excised products between 200 and 500 nucleotides from a 1% agarose gel and purified them with the Qiagen gel purification kit. We analyzed the libraries on the Bioanalyzer 2100 and then combined the products in equimolar ratios from the eight different degenerate primer reactions. We sent the libraries for next-generation sequencing on the Illumina HiSeq 2500, aligned resulting reads with Bowtie2, and sorted the reads based on the presence or absence of L1 sequence. We also segregated and excluded all previously published polymorphic L1 insertions and reference insertions from our list of putative somatic insertions using the L1-seq pipeline established by Adam Ewing (249). Overall, our bioinformatics analysis was identical to previous studies (206,249).

Stringency analysis.

We filtered the putative somatic insertions detected by L1-seq using a map score of 0.5 or greater, a nucleotide window of 90 base pairs or greater, and 3 or more unique reads. After filtering with the aforementioned criteria, we then attempted our PCR validations with both

conventional and nested PCR in all samples. There parameters gave relatively good validation rates of 50% and ~69% for the two groups of SCC samples respectively and are comparable to validation rates in previous publications (203,204,206).

Immunohistochemistry.

We performed our immunohistochemistry (IHC) experiments using the EnVision System-HRP (catalog K4006; Dako) according the manufacturer's protocol. We performed the primary antibody incubation with a mouse monoclonal ORF1 (1.25 mg/mL) at a 1:3,000 dilution for 40 minutes at room temperature. We performed secondary antibody incubation as per the manufacturer's protocol. The monoclonal antibody used detects amino acids 35-44 of the ORF1 protein and is the same antibody we used previously (206).

Results

L1 Protein Expression in NE and SCC in the Esophagus

In order to establish L1 activity in patients with SCC and the normal esophagus of SCC patients, we evaluated L1 protein expression in both normal and tumor samples. Although LINE-1 protein expression is necessary for LINE1 mobilization (282), it is not sufficient to ensure somatic insertions will occur, even in cancer (Doucet-O'Hare et al., 2015). Rodić and colleagues suggested that ORF1p expression is a potential biomarker in cancer due to its pervasive expression in many types of epithelial cancer (209). Even in the absence of confirmed somatic insertions, we previously observed ORF1p expression in all esophageal adenocarcinoma samples evaluated(206). Furthermore, we observed weaker ORF1p expression in the normal squamous epithelial tissues of many patients coinciding with a few insertions that likely occurred in the normal tissue originally and expanded in the resulting lesion (145).

Due to the expression of ORF1p in many epithelial cancers (208), we hypothesized there would be ORF1p expression present in SCC. Additionally, we expected to observe weaker ORF1p expression in the normal squamous epithelium of the patients (145). We obtained formalin-fixed paraffin-embedded tissue samples for four of the nine patients, and using a monoclonal OFF1p antibody (256) confirmed ORF1p expression in all samples evaluated. Three of the four cases evaluated with IHC showed robust ORF1p staining; however, in one case it was much weaker (Fig. 3.1). Intriguingly, the SCC case in which ORF1p expression was low was the only case in which no somatic insertions were confirmed (Table 3.2). We also observed low-level ORF1p expression in normal squamous epithelium of all patients evaluated (Fig. 3.1).

ORF1p expression in epithelial cancer typically appears in a diffuse to speckled distribution in the cytoplasm of the cancer cells. Although ORF1p was present in all four samples evaluated, the distribution pattern differed among samples and across regions of the same sample. In some samples, we observed ORF1p expression predominantly in a diffuse pattern in the cytoplasm of the cells, while in three samples, we observed an accentuated perinuclear pattern oftentimes with aggregates of the protein localizing near the nuclear periphery (Fig. 3.2). We confirmed the perinuclear staining pattern with a second antibody targeting a different portion of the ORF1 protein (206); however, the significance of these patterns is still unknown.

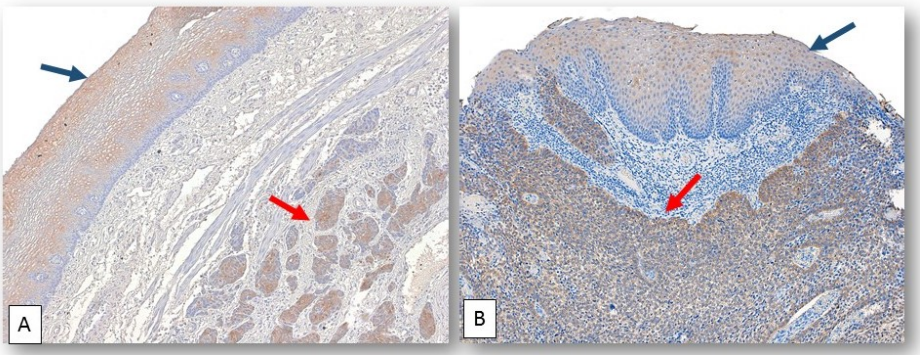


Figure 3.1: ORF1p expression in normal esophagus and squamous cell carcinoma A) and B) Representative photomicrographs depicting LINE-1 ORF1p expression in normal esophageal squamous epithelium (black arrows) and in esophageal squamous cell carcinoma cases (red arrows): A) and B) Esophageal tissue with normal squamous epithelium adjacent to squamous cell carcinoma from two separate individuals stained with LINE1 ORF1p (final magnification x100).

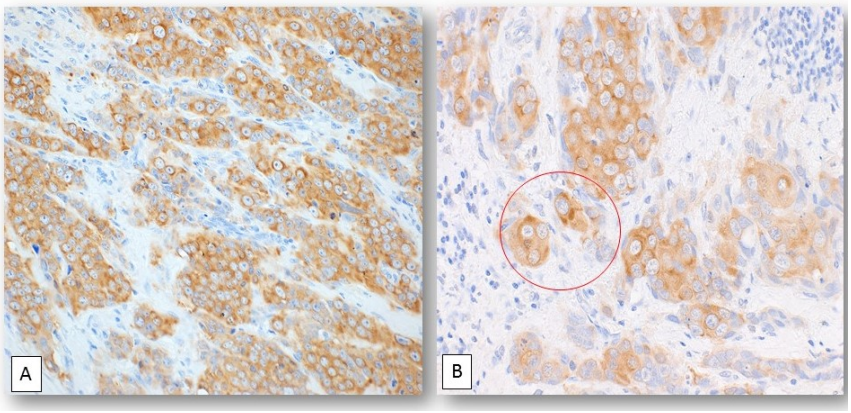


Figure 3.2: ORF1p expression patterns in esophageal squamous cell carcinoma

A) and B) photomicrographs showing an invasive squamous cell carcinoma case where peri-nuclear staining patterns manifest for ORF1p. A) Final magnification x100. B) The same case as A) with a higher magnification view of ORF1p peri-nuclear staining accentuation (within red circle). Final magnification x160.

Esophageal Squamous Cell Cancer Validated Insertions														
Insertion number	Predicted breakpoint	Group	Sample	Patient Age	5' validated?	Inversion?	TSD sequence	Insertion size (w/o poly A)	Exact Genomic Breakpoint	In normal w/ nested	Amplified with single PCR	Endo site	Gene	Strand of insertion
1	chr1:74225746	2	20T	79	Yes	No	300 bp	120 bp	chr1:74225531		Yes	AGAT/AA		"-"
2	chr1:91054960	2	20T	79										"+"
3	chr1:110159225	2	20T	79							Yes			"+"
4	chr1:188390840	2	21T	71										"+"
5	chr2:7391894	2	23T	76						Yes				"-"
6	chr3:106586376	2	20T	79									LINC00882	"-"
7	chr3:158778615	2	21T	71	Yes	Yes	15 bp	1859 bp	chr3:158778722		Yes	TTCT/GA		"-"
8	chr3:173244207	2	21T	71							Yes		NLGN1	"+"
9	chr4:21015215	2	21T	71	Yes	Yes	11 bp	1229 bp	chr4:21015270		Yes	CTTT/AT	KCNIP4	"-"
10	chr4:21295696	2	20T	79	Yes	No	16 bp	520 bp	chr4:21295716			TATG/AA	KCNIP4	"+"
11	chr4:33062493	2	20T	79										"-"
12	chr4:34773929	2	23T	76							Yes			"-"
13	chr4:100470748	2	20T	79									TRMT1	"-"
14	chr4:122022647	2	20T	79	Yes	No	5 bp	1110 bp	chr4:122022858			AAAA/CA		"-"
15	chr5:18343101	2	20T	79							Yes			"+"
16	chr5:22972054	2	21T	71							Yes			"+"
17	chr5:27118145	2	23T	76	Yes	No	376 bp	121 bp	chr5:27117892		Yes	TATA/CA		"-"
18	chr5:63570580	2	21T	71	Yes	No	389 bp deletion	630 bp	chr5:63571024		Yes	Endo indep	RNF180	"-"

Insertion number	Predicted breakpoint	Group	Sample	Patient Age	5' validated?	Inversion?	TSD sequence	Insertion size (w/o poly A)	Exact Genomic Breakpoint	In normal w/ nested	Amplified with single PCR	Endo site	Gene	Strand of insertion
19	chr5:146263689	2	21T	71	Yes	No	12 bp	1039 bp	chr5:146263701	Yes	Yes	ACCA/GC	PPP2R2B	"+"
20	chr6:54007315	2	20T	79	Yes	Yes	12 bp	258 bp	chr6:54007426			AAAG/AC	MLIP	"-"
21	chr6:65430273	2	20T	79						Yes	Yes		EYS	"-"
22	chr6:93175977	2	20T	79							Yes			"+"
23	chr6:149512094	2	21T	71										"+"
24	chr6:167049997	2	21T	71									RPS6KA2	"-"
25	chr7:78377347	2	20T	79	Yes	No	20 bp	773 bp	chr7:78377476		Yes	TTTT/TT	MAGI2	"-"
26	chr7:84066697	2	22T	59	Yes	No	1 bp deletion	354 bp	chr7:84066755	Yes	Yes	Endo indep		"-"
27	chr7:113098298	2	22T	59	Yes	No	None	1189 bp	chr7:113098416	Yes	Yes	GTTA/TA		"-"
28	chr7:144876298	2	21T	71	Yes	No	11 bp	361 bp	chr7: 144876634			GCAG/TG		"-"
29	chr8:62731519	2	20T	79	Yes	No	5 bp	568 bp	chr8:62731576			TAAA/AA		"-"
30	chr8:63452731	2	23T	76						Yes	Yes		NKAIN3	"-"
31	chr8:78650641	2	23T	76						Yes	Yes			"+"
32	chr8:96814193	2	20T	79	Yes	No	2 bp	737 bp	chr8:96814250	Yes	Yes	CTTG/AA	C8orf37-AS1, LOC100616530	"-"
33	chr8:105742587	2	21T	71										"+"
34	chr8:130050500	2	23T	76						Yes	Yes			"+"

Insertion number	Predicted breakpoint	Group	Sample	Patient Age	5' validated?	Inversion?	TSD sequence	Insertion size (w/o poly A)	Exact Genomic Breakpoint	In normal w/ nested	Amplified with single PCR	Endo site	Gene	Strand of insertion
35	chr8:137876889	2	21T	71										"-"
36	chr8:138433870	2	21T	71										"+"
37	chr9:27521686	2	21T	71									MOB3B	"+"
38	chr9:30914267	2	21T	71	Yes	Yes	14 bp	1305 bp	chr9:82494253		Yes	ATCA/AG		"+"
39	chr9:82494252	2	21T	71	Yes	No	12 bp	282 bp	chr9:30914249			ATCA/AG	LINC01507	"+"
40	chr10:11187849	2	22T	59							Yes		CELF2	"+"
41	chr10:55846139	2	23T	76	Yes	No	3 bp deletion	175 bp	chr10:55846282	Yes	Yes	Endo indep	PCDH15	"-"
42	chr10:115267083	2	23T	76										"+"
43	chr11:23140915	2	20T	79	Yes	No	14 bp	1043 bp	chr11:23140929			AACA/TC		"+"
44	chr11:43634948	2	23T	76										"+"
45	chr11:78577470	2	20T	79	Yes	Yes	109 bp deletion	346 bp	chr11:78577397		Yes	Endo indep	TENM4	"+"
46	chr11:95617289	2	23T	76						Yes	Yes		MTMR2	"-"
47	chr12:72582208	2	21T	71						Yes	Yes			"+"
48	chr13:90364217	2	20T	79										"-"
49	chr14:21603456	2	20T	79							Yes			"+"
50	chr14:95721343	2	20T	79									CLMN	"+"
51	chr15:55401145	2	20T	79										"-"
52	chr16:32621143	2	23T	76										"-"

Insertion number	Predicted breakpoint	Group	Sample	Patient Age	5' validated?	Inversion?	TSD sequence	Insertion size (w/o poly A)	Exact Genomic Breakpoint	In normal w/ nested	Amplified with single PCR	Endo site	Gene	Strand of insertion
53	chr18:7914950	2	21T	71									PTPRM	"+"
54	chr19:31676750	2	23T	76										"-"
55	chr20:12193544	2	21T	71										"+"
56	chr20:31767412	2	20T	79									BPIFA2	"+"
57	chr4:178589680	1	2T	67							Yes		AK094945	"-"
58	chr5:8875319	1	3T	51							Yes		BC032891	"+"
59	chr14:81244336	1	3T	51							Yes		CEP128	"-"
60	chr1:169592527	1	3T	51	Yes	No	3 bp (	290 bp	chr1:169592905		Yes	GATT/AA	SELP	"-"
61	chr2:62784947	1	3T	51							Yes			"-"
62	chr2:118147901	1	1T	76							Yes			"+"
63	chr2:192054832	1	1T	76							Yes			"-"
64	chr5:6447510	1	8T	78							Yes			"-"
65	chr7:69594334	1	8T	78	Yes	No	2 bp	365 bp	chr11:122340070		Yes	CAGT/AT	AUTS2	"+"
66	chr11:122339699	1	3T	51							Yes			"-"
67	chr16:20172684	1	3T	51							Yes			"-"
68	chrX:144755675	1	3T	51							Yes			"-"

Insertion number	Predicted breakpoint	Group	Sample	Patient Age	5' validated?	Inversion?	TSD sequence	Insertion size (w/o poly A)	Exact Genomic Breakpoint	In normal w/ nested	Amplified with single PCR	Endo site	Gene	Strand of insertion
69	chr3:108723582	1	8T	78							Yes		MORC1	"+"
70	chr6:5599305	1	3T	51									FARS2	"-"
71	chr7:18316210	1	8T	78									HDAC9	"-"
72	chr7:71682380	1	8T	78	Yes	No	6 bp	528 bp	chr7:71682386			AAAA/CA	CALN1	"+"
73	chr9:8381705	1	1T	76									PTPRD	"+"
74	chr9:81288257	1	8T	78										"+"

Table 3.2: This table lists all confirmed somatic insertions, the sample in which the insertion occurred in addition to all pertinent characteristic of the insertion such as endonuclease cleavage site, poly (A) tail length, genomic breakpoint, the gene into which the insertion occurred, the strand into which the insertion occurred, etc.

Somatic L1 Insertions in Squamous Cell Carcinoma of the Esophagus

After observing ORF1p expression in all SCC cases evaluated, we characterized the potential mutations caused by L1 activity in esophageal squamous cell carcinoma by studying the same patients' matched fresh-frozen normal and cancer tissues. Because SCC is a relatively rare cancer in the United States, we were only able to acquire nine samples total for our study (283). We received the samples in two groups, the first group consisted of four individuals with SCC and one individual with esophageal adenocarcinoma (EAC), and the second group consisted of five individuals with SCC (Table 3.1). We prepared DNA from each of the two groups into L1-seq libraries separately and the samples were next-generation sequenced and analyzed. L1 seq is a high throughput technique which enriches for the human-specific sub-family of L1 elements using specific PCR primers.(249). After sequencing, the data were subjected to a computational pipeline, designed by Ewing and colleagues, that analyzes sequencing data and identifies potential somatic insertions present in tissue

Evidence of active L1 elements manifests in the form of both protein expression and "somatic" insertions. As in our previous studies, we defined somatic insertions as those which were not inherited from a previous generation and therefore present in only a subset of cells within a tissue (206). Although the majority of somatic insertions are expected to occur in the tumor and to be absent from the normal tissue, we hypothesized that some somatic insertions may exist in a sub-clonal population of normal esophageal cells. The sub-clonal insertions could be amplified when cells are selectively amplified during tumor initiation and progression (206). Due to the previously observed sub-clonal insertions in normal esophagus and Barrett's esophagus (206), we evaluated every putative insertion with both conventional PCR and nested PCR in both normal and tumor DNA (Fig. 3.3 and 3.4).

In order to select putative insertions for validation with PCR and Sanger sequencing, we filtered our results by selecting insertions with 3 unique reads or more, greater than a 90 base-pair nucleotide window, and a map score of 0.5 or greater (249). After filtering in the first group of samples, we observed 100 potential tumor only insertions and tested 36 of them. Using PCR and Sanger sequencing we were able to validate 18 insertions distributed among the four individuals with SCC. We did not validate any somatic insertions in the individual with EAC; however, we previously validated somatic insertions in other patients with EAC (206).

After filtering the data for the second group of samples, there were 133 potential insertions unique to the tumor, 82 of which were tested. We validated 56 insertions distributed among four of the five individuals with 12 of the insertions appearing to be sub-clonal in the normal esophagus and clonal in the tumor (Fig. 3.3). Interestingly, all four patients with confirmed somatic insertions in the second group of samples harbored sub-clonal insertions in the normal tissue. We previously detected somatic insertions in normal tissue with nested PCR for approximately 5% of somatic insertions (204,206). In esophageal squamous cell carcinoma samples, we observed that ~16% (12/74 insertions) were present in the normal esophagus with nested PCR (Figures 3.3 and 3.4). We speculate that the sub-clonal insertions in the normal esophagus could have occurred in a single cell which then transdifferentiated into dysplasia and then squamous cell carcinoma. In the process of tumorigenesis, the cells with the insertions may have been selectively amplified and thus the insertions in the tumor cells were easily detectable with a conventional PCR (206,264). Furthermore, the incidence of sub-clonal insertions in the normal tissue does not appear to be a rare phenomenon in SCC as it is present in four out of the eight patients with confirmed somatic insertions (Table 3.2). The sub-clonal insertions observed

in the normal tissue are unlikely to be a result of contamination from the tumor DNA because each of these samples was micro-dissected under the guidance of a gastrointestinal pathologist.

Unfortunately, only normal esophagus and tumor were available from the patients; therefore, we cannot be certain that the confirmed somatic insertions observed in both tissues were not low-level germline insertions. The sensitivity of the conventional PCR method for somatic insertion confirmation is approximately one in ten, e.g. if one in ten cells has the insertion present it is consistently detected at a low level by conventional PCR(204). Using nested PCR, we consistently detect an insertion present in one out of a thousand cells (204). Although it is unlikely for a sub-clonal insertion to be germline, we cannot determine that the sub-clonal insertions detected in normal tissue are definitive somatic insertions. To date, there are no known germline insertions present at such a low frequency in an individual's tissue. Tumor-specific insertions detected are also unlikely to be germline because they were absent from normal esophageal tissue with nested PCR (Fig 3.5).

Characterization of Tumor Specific Insertions

After establishing that L1 was actively mobile in eight of the nine cases of SCC evaluated, we evaluated more thoroughly the characteristics of the confirmed somatic insertions. To identify the precise base pair at which the insertions occurred we performed PCRs to amplify the 5' end of the insertions and were able to successfully amplify and sequence 23/74 (Table 3.2). For a subset of confirmed insertions, we identified endonuclease cleavage sites and target site duplications (TSDs) which are both hallmarks of the process of retrotransposition (Table 3.2). Out of the 20 putative endonuclease cleavage sites identified in our study, 9 were similar (differed by 3 base pairs or less) to the canonical endonuclease cleavage site in target-primed reverse transcription reactions (111). With regard to 4 insertions presumed to be endonuclease

independent (284), all had deletions at the site of insertion integration (265). For some of the confirmed insertions, we identified the TSDs ranging from two base pairs in length to 389 base pairs in our samples. The insertion sizes varied from 120 to 1,859 base pairs with 11 of the insertions under 500 base pairs (Table 3.2). In a previous study, we observed similar results with regard to insertion size, endonuclease cleavage sites, and insertion sizes (206). Furthermore, five of the insertions had 5' end inversions, a finding consistent with previous studies of germ-line insertions and L1 insertions in cancer (111,201–203,205,206,210). Of the insertions validated in SCC patients, 40 amplified easily with a single PCR, while the remaining 34 insertions amplified only following a nested PCR (Fig. 3.3). We confirmed 12 insertions in normal tissue with only a nested PCR, but those insertions amplified with a conventional PCR in the tumor DNA (Fig 3.4).

At least half of the somatic insertions validated in the SCC patients had characteristics that differ from a canonical germline insertion (51,111). We confirmed that 4/23 insertions had deletions (17.4%) at the site of integration; however, only 10% of germline insertions normally have integration-site deletions (111). The four insertions with integration site deletions also lacked clear target-site duplications, both of which are clear hallmarks for canonical retrotransposition, suggesting they were endonuclease independent (254). Furthermore, a little less than half of all validated somatic insertions were 500 base pairs or less due to severe 5' truncation. In contrast to others' work, we did not detect any 3' transductions (202); however, L1-seq rarely detects such events.

None of the somatic insertions evaluated occurred in exons; however, 32 insertions did occur into the introns of 32 different genes. One of the insertions occurred into an intron of two different genes, namely C8orf37-AS1 and LOC100616530. Another gene, KCNIP4, had two different confirmed insertions in an intron approximately 280 kb apart (Fig 3.6). Although

KCNIP4 is approximately 1.2 Mb in size, the insertions occurred in two different individuals, meaning that this gene was recurrently subjected to L1 insertions. The insertions in KCNIP4 are in opposite orientations relative to the gene. Of the 32 genes with validated somatic insertions, 22 occurred into genes that have previously been associated with cancer (Table 3.3). Our findings reveal a statistically significant enrichment of validated somatic insertions into genes previously associated with cancer ($P < 1 \times 10^{-10}$) (284-327). We considered the probability that a somatic insertion would hit more of the cancer-associated genes than would be expected due to chance alone. After accounting for gene size, we observed a significant enrichment of somatic insertions into cancer associated genes using a chi squared test ($P < 1 \times 10^{-10}$). Four of the 22 genes are considered "known cancer genes" by the network of cancer genes (285), and the remaining 18 genes are candidate cancer genes which have been associated with cancer in the literature (Table 3.3)(286–328). Interestingly, out of the 22 genes associated with cancer, sequence variation in 16 of the genes have also been associated with smoking (Table 3.3)(329–336). Smoking is one of the main risk factors for SCC; therefore, it is fitting that so many of the cancer-associated genes into which L1 elements inserted are associated with smoking (Table 3.3).

Gene Name(s)	Cancer type(s)	NGC 5.0	Associated with smoking?	Cancer References	Smoking References	Entrez Gene
EYS	esophageal, lung	Y	Y	286, 287	329. 330	346007
PCDH15	head and neck, pancreatic	Y	Y	288, 289	330, 331	65217
HDAC9	adrenocortical adenoma, squamous cell carcinoma	Y	Y	290, 291	330	9734
PTPRD	lung, breast, pancreas, gastric adenocarcinoma	Y	Y	292, 293, 294, 296	330, 331, 332	5789
KCNIP4	renal cell carcinoma, haematological	N	Y	295, 297	330, 331, 332	80333
RNF180	gastric	N	Y	298, 299	330, 331, 333	285671
PPP2R2B	breast, colorectal	N	Y	300, 301		5521
TRMT1	prostate	N		302	330, 331, 334, 335	55621
RPS6KA2	ovarian, colorectal, breast, pancreatic	N	Y	303, 304, 305, 306		6196
C8ORF37-AS1	colorectal	N		307	330, 336	100616530
MOB3B	prostate	N	Y	308	330, 331	79817
CELF2	breast, colon, sqaumous cell carcinoma	N	Y	309, 310, 311	330	10659
MAGI2	lung adenocarcinoma	N		312		9863
PTPRM	breast, glioblastoma, B cell leukemia	N	Y	313, 314, 315	330	5797
BPIFA2	salivary gland tumors	N		316, 317		140683
SELP	colon caricnoma metastasis, colorectal cancer	N		318, 319, 320		6403
AUTS2	haematological	N	Y	297	330	26053
MORC1	colorectal cancer	N	Y	321	330	27136
NLGN1	oral squamous cell carcinoma, uterine leiomyosarcoma	N	Y	322, 323	330	22871
FARS2	colon, colorectal	N	Y	324	330	10667
CALN1	gastric	N	Y	325	330	83698
TENM4/ODZ4/DOC4	primary lymphoma of CNS, neuroblastoma, breast cancer	N	Y	326, 327, 328	330	26011

Table 3.3: The gene listed in this table all had validated somatic L1 insertions into introns. All of these genes have been associated with cancer and many of them have also been associated with smoking, a major risk factor of esophageal squamous cell carcinoma.

Discussion

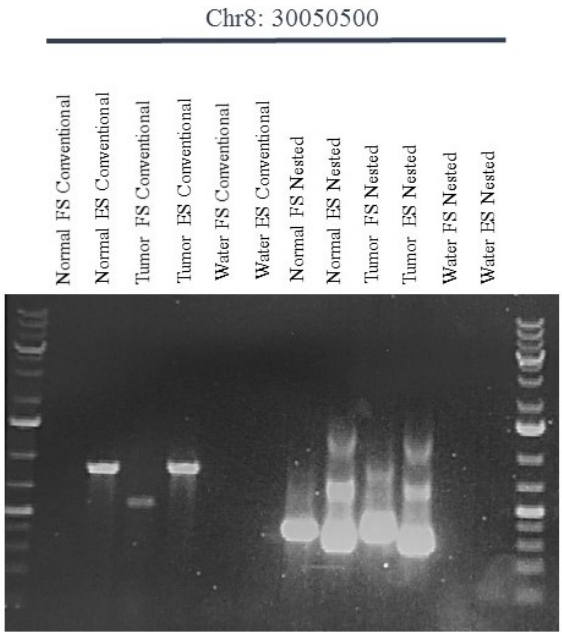
Expanding our understanding of carcinogenic mutational mechanisms will lead to earlier diagnosis and better treatment opportunities for cancer patients. Although esophageal squamous cell carcinoma is rare in the United States, its prevalence throughout the rest of the world, necessitates its study. In order to detect, diagnose, and treat SCC effectively, we need a thorough evaluation of mutations acquired in the squamous tissue of the esophagus, both in the normal tissue and in dysplastic tissue, which transitions to cancer. Cellular processes often become dysregulated during carcinogenesis and may provide a favorable environment for the activity of retrotransposons.

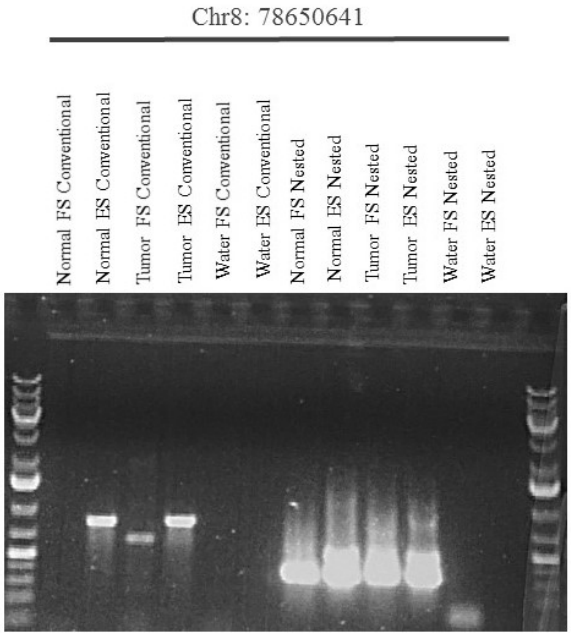
Before evaluating cancer samples for newly acquired somatic insertions, we looked for L1 protein expression differences between the normal squamous epithelium and the tumor. We observed higher ORF1p expression in the SCC patients possessing a larger number of confirmed somatic insertions. Because we were only able to acquire tissue for four of the nine SCC patients studied, our results are not statistically significant. Previously, we hypothesized that higher ORF1p expression correlates with a higher number of somatic insertion occurrences in patients; nevertheless, this did not hold true for our EAC and Barrett's esophagus patients. Interestingly, in the present study multiple patterns of ORF1p staining appeared in the cancer of three individuals including a diffuse cytoplasmic pattern and a perinuclear pattern with accentuation and protein aggregates near the nuclear periphery (Fig. 3.2). While we do not fully understand the significance of these ORF1 protein expression patterns in SCC, we can conclude that the protein is present in all samples evaluated, and that there are differences in level and pattern of expression between patient samples.

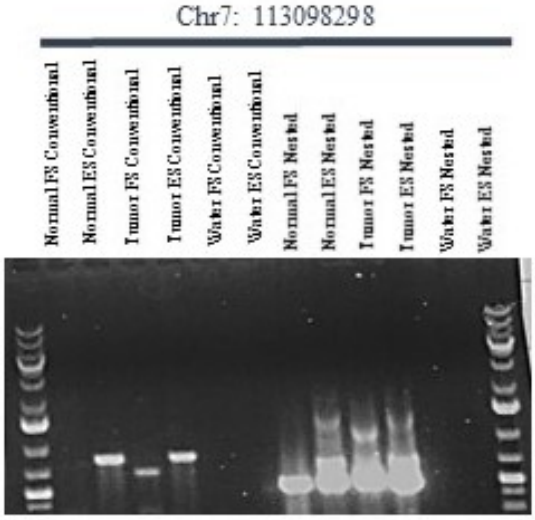
Our group and many others have firmly established that L1 somatic insertions occur frequently in epithelial cancers and presented evidence that insertions may sometimes contribute to cancer development (201–206,210,237). In this study, we demonstrate retrotransposition is an active process in eight out of nine patients with SCC. We observed 62 somatic insertions absent from normal tissue and present in the tumor and 12 insertions that amplified easily with a conventional PCR in tumor, but were also present in the matched normal tissue using nested PCR (Fig. 3.3, 3.4). There was no significant correlation between age and insertion occurrence (Fig. 3.7).

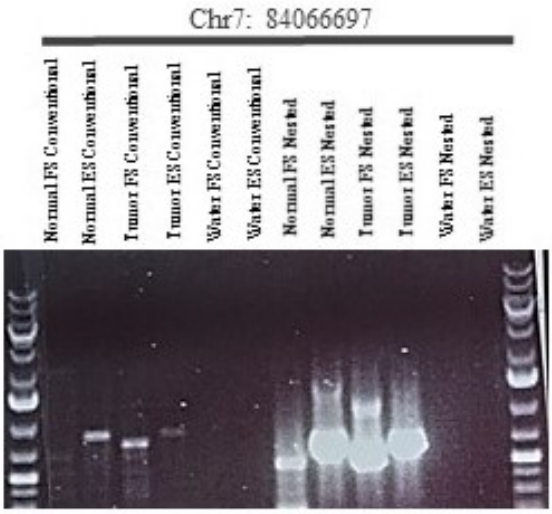
Faulkner and colleagues first detected somatic insertions into normal hippocampus and a few years later, we detected a somatic insertion in normal colon and two somatic insertions in normal esophagus (204,206,220). Aside from the aforementioned examples, observations of somatic insertions in normal tissues are relatively uncommon (221). Interestingly, we detected twelve instances of somatic insertions in normal tissue with nested PCR suggesting the insertions are in only a few cells. With conventional PCR, the same insertions are observable in tumor DNA exemplifying that a larger number of cells contain the insertions in the tumor. Because the frequency of sub-clonal insertions in normal esophagus of SCC patients is much higher than we have previously observed in cancer patients, it appears that the esophagus is permissive for L1 somatic insertions (204,206). Our data are consistent with two different models explaining the role of retrotransposition in cancer. First, it is possible, as previously suggested(204,206,264), that some somatic L1 insertions are acquired in the normal tissue and subsequently expand in the cancer (Fig. 3.5). In future studies, single-cell sequencing should reveal the occurrence of somatic L1 insertions in normal cells and give insight with regard to their frequency.

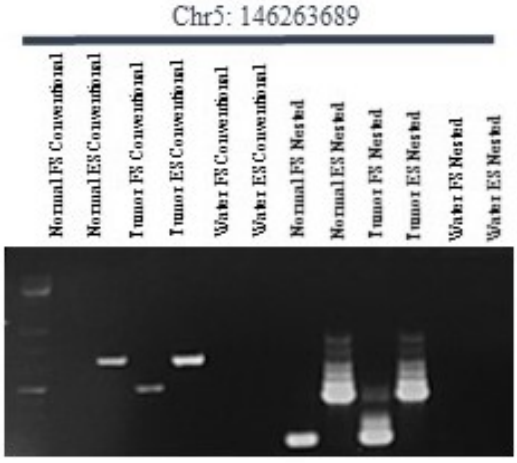
Many groups detected somatic retrotransposition events in various types of epithelial cancer such as squamous cell lung, head and neck, hepatocellular, breast, prostate, colorectal, bone, esophageal adenocarcinoma, pancreatic ductal carcinoma, and endometrial cancer (202–206,208,210,337). We confirmed somatic retrotransposition is active in esophageal squamous cell carcinoma in eight of nine patients evaluated; furthermore, we observed twelve instances where retrotransposons actively mobilized in normal squamous epithelium of the esophagus. Understanding the role of somatic insertions in cancer relies heavily on having an accurate estimation of retrotransposon activity in normal cells. In light of these data, we must establish the underlying rate of retrotransposition in any normal tissue subsequently evaluated in a disease state in order to interpret the data obtained.


Chr8: 30050500
Normal FS Conventional
Normal ES Conventional
Tumor FS Conventional
Tumor ES Conventional
Water FS Conventional
Water ES Conventional
Normal FS Nested
Normal ES Nested
Tumor FS Nested
Tumor ES Nested
Water FS Nested
Water ES Nested


Chr8: 78650641
Normal FS Conventional
Normal ES Conventional
Tumor FS Conventional
Tumor ES Conventional
Water FS Conventional
Water ES Conventional
Normal FS Nested
Normal ES Nested
Tumor FS Nested
Tumor ES Nested
Water FS Nested
Water ES Nested


Chr7: 113098298
Normal FS Conventional
Normal ES Conventional
Tumor FS Conventional
Tumor ES Conventional
Water FS Conventional
Water ES Conventional
Normal FS Nested
Normal ES Nested
Tumor FS Nested
Tumor ES Nested
Water FS Nested
Water ES Nested


Chr7: 84066697
Normal FS Conventional
Normal ES Conventional
Tumor FS Conventional
Tumor ES Conventional
Water FS Conventional
Water ES Conventional
Normal FS Nested
Normal ES Nested
Tumor FS Nested
Tumor ES Nested
Water FS Nested
Water ES Nested


Chr5: 146263689
Normal FS Conventional
Normal ES Conventional
Tumor FS Conventional
Tumor ES Conventional
Water FS Conventional
Water ES Conventional
Normal FS Nested
Normal ES Nested
Tumor FS Nested
Tumor ES Nested
Water FS Nested
Water ES Nested

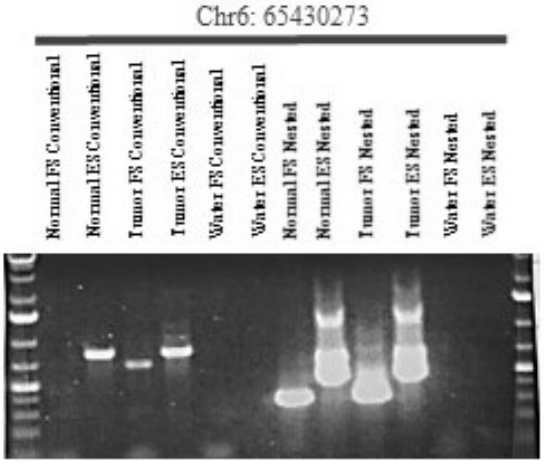

Chr6: 65430273
Normal FS Conventional
Normal ES Conventional
Tumor FS Conventional
Tumor ES Conventional
Water FS Conventional
Water ES Conventional
Normal FS Nested
Normal ES Nested
Tumor FS Nested
Tumor ES Nested
Water FS Nested
Water ES Nested

Figure 3.3: Examples of sub-clonal insertions in normal esophagus Conventional PCRs done on normal esophagus and esophageal squamous cell carcinoma alongside the corresponding nested PCRs. The nested PCRs use the product from the conventional PCRs and a new pair of primers, which are nearer to the predicted breakpoint of the somatic insertion. These PCRs showcase events which presumably occur in the normal esophageal tissue and are expanded clonally in the resulting tumors.

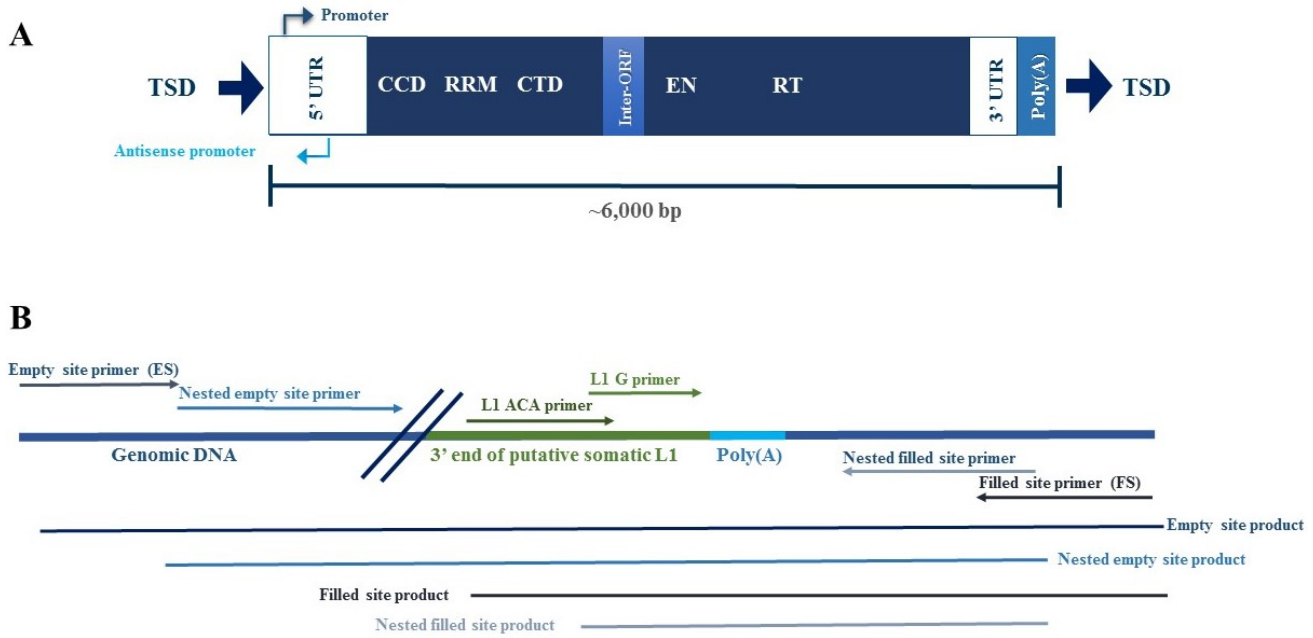


Figure 3.4: L1 structure and L1-seq validation scheme A) LINE-1 structure of a human specific active element in the human genome. Full-length L1 elements are approximately 6,000 nucleotides in length and have a 5' untranslated region, containing both a promoter and an anti-sense promoter. L1 encodes two open reading frames for proteins on which it relies for its mobilization in the cell (21). The first open reading frame encodes ORF1p, a protein with RNA chaperoning activity that includes a coiled-coil domain (CCD), an RNA recognition motif (RRM), and a carboxy-terminal domain (CTD) (24,29). The second open reading encodes ORF2p, possesses an endonuclease domain (EN) and a reverse transcriptase domain (RT) (25,36,38). The element also has a poly(A) tail between the 3' untranslated region and the target site duplication (TSD). Both the poly(A) tail and TSDs are hallmarks of the target-primed reverse transcription, the process by which L1 elements mobilize (49). B) The PCR validation method for predicted somatic insertions utilizes two unique areas of sequence in the 3' end of the L1 element. The sequence 'ACA' is incorporated into the L1 ACA primer and is positioned

about 90 nucleotides from the poly(A) tail. The 'ACA' nucleotides (94,95) distinguish the human specific transcriptionally active L1 element from other L1s in the genome and allow for a specific PCR product to amplify. The L1 G primer has a 'G' nucleotide which is also unique to the human specific active L1 element and the nucleotide is approximately ten nucleotides proceeding the poly(A) tail (249). In order to validate an insertion, pictured adjacent to a poly-A tail, the L1 primers are used in conjunction with filled and empty site primers to amplify the 3' end of the somatic insertion. The nested empty site and nested filled site primers are flanked by the empty and filled site primers and are used to verify that insertions predicted in tumor are truly absent from normal tissue. Additionally, nested primers are utilized when amplifying low copy number somatic insertions because they are able to detect when one cell out of a thousand has a copy of an insertion. The size of the expected products and relative locations of the primers are pictured in the figure.

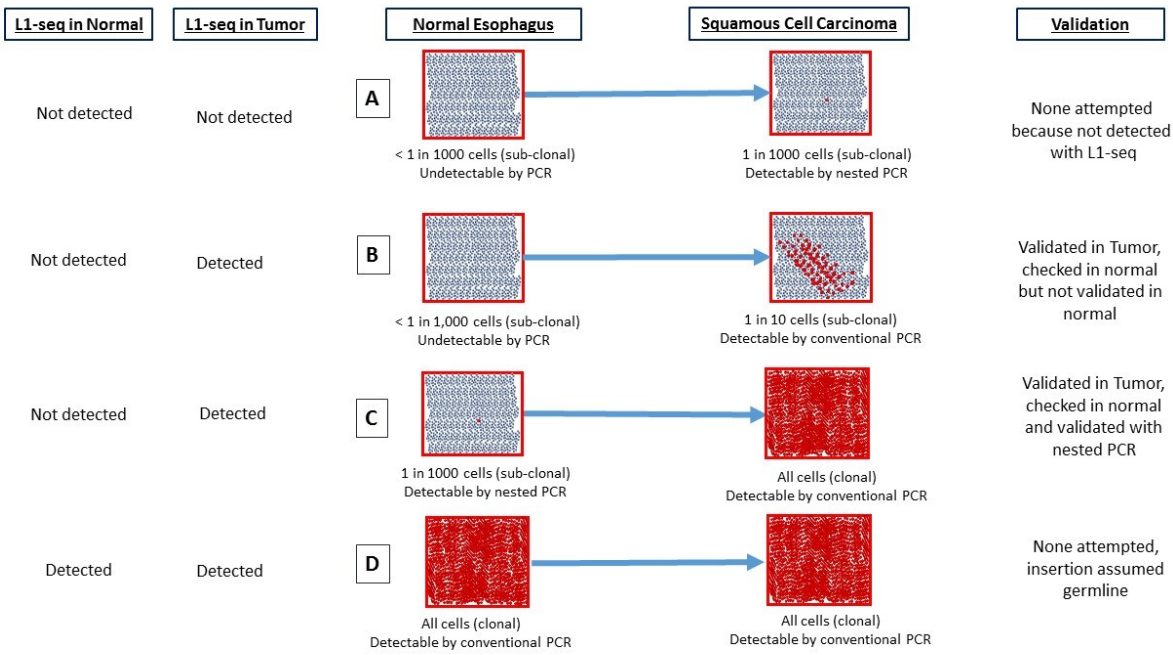


Figure 3.5: Acquisition, Detection, and Validation of sub-clonal and clonal somatic insertions in L1-seq This diagram details the sensitivity of L1-seq with regard to detecting somatic insertions at differing levels of clonality in a tissue and different scenarios by which a somatic insertion could become amplified in a tumor. (A) An insertion at a very low frequency (less than one in a thousand cells) in the normal esophagus and at an undetectable level in the tumor when evaluated by L1-seq. (B) An insertion at an undetectable level in the normal esophagus and a sub-clonal but detectable level in the tumor. (C) An insertion at an undetectable level for L1-seq in the normal esophagus which subsequently becomes clonal in the tumor. The insertion is not detectable with conventional PCR in the normal, but can be detected with a nested PCR. (D) An insertion which is clonal in both the normal esophagus and the tumor. This insertion is not tested by PCR because it is presumed to be germline.

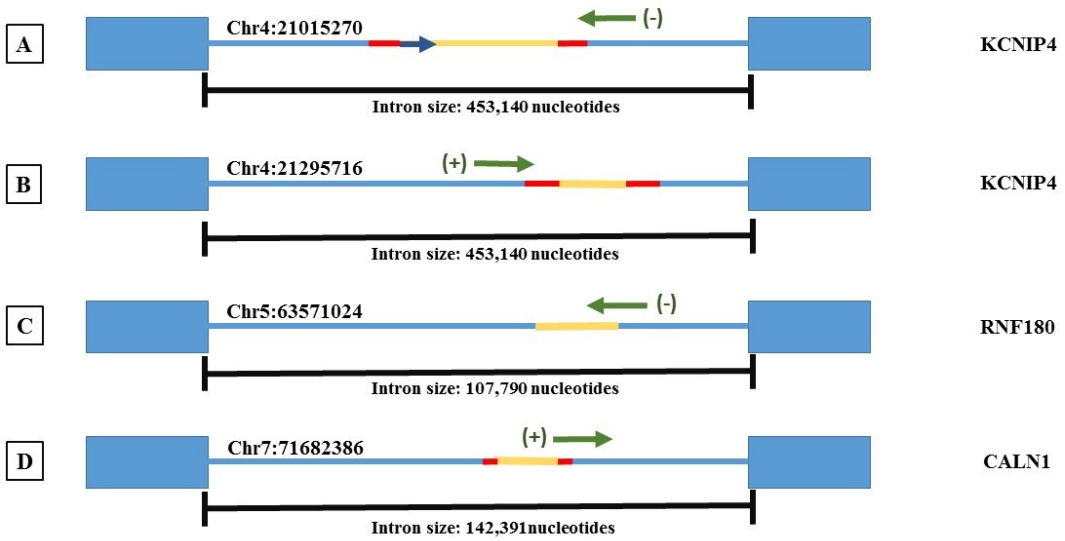


Figure 3.6: Somatic insertions occurrence, ORF1p expression, and patient age. This figure details the ages of the patients compared with the number of somatic insertions which occurred in each individual and the level of ORF1p expression observed in the individual. The graph shows there is a small positive correlation between the age of the patients and the insertion occurrence into them, e.g. older patients tend to have more somatic insertions. However, the positive correlation between age and insertion occurrence is not statistically significant when a linear regression is performed ($P = 0.1025$).

Somatic insertion occurrence, ORF1p expression, and patient age			
Sample designation	Age	insertions	Expression
1N/T	76	3	N/A
2N/T	67	1	N/A
3N/T	51	3	N/A
8N/T	78	6	N/A
20N/T	79	22	+++
21N/T	71	19	+++
22N/T	59	3	N/A
23N/T	76	13	+++
24N/T	58	0	+

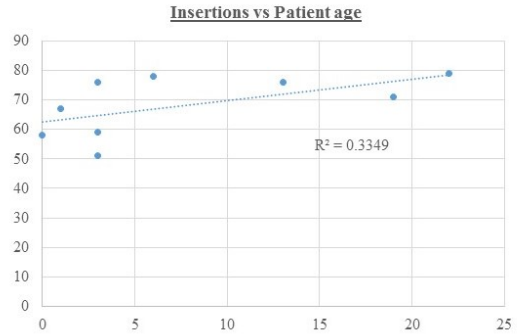


Figure 3.7: Somatic insertions occurrence, ORF1p expression, and patient age.

This figure details the ages of the patients compared with the number of somatic insertions which occurred in each individual and the level of ORF1p expression observed in the individual. The graph shows there is a small positive correlation between the age of the patients and the insertion occurrence into them, e.g. older patients tend to have more somatic insertions. However, the positive correlation between age and insertion occurrence is not statistically significant when a linear regression is performed ($P = 0.1025$).

CHAPTER 4: Future Directions

Retrotransposition: Cause or Consequence?

Even after all the work done to characterize mobile DNA activity in various tissues and diseases we are still unable to definitively state that retrotransposons are more than a rare cause of disease. Although there are more than one hundred examples of disease due to retrotransposons, it is certainly more of an exception than a rule. Many, including myself, have argued that due to the dysregulation of many cellular processes during carcinogenesis and thereafter, retrotransposons may simply be less restricted and therefore successfully mobilize at a higher rate than in normal tissues. Evidence of active retrotransposition has been observed in many cancers (202–206,210,244,337) and is notably frequent in epithelial cancers.

Although many groups have reported evidence of somatic insertions in cancer, very few have shown evidence that a somatic insertion contributed to the development of cancer. A somatic insertion into the *APC* gene of a patient with colorectal cancer was the first observed example of a L1 insertion contributing to carcinogenesis in 1992 (237). Many years later, another definitive somatic insertion was found in an exon of the *PTEN* tumor suppressor gene (210). Finally, a tumor-specific, presumably somatic, insertion into the gene ST18 was observed in a patient with hepatocellular carcinoma (205). The insertion activated the suppression of tumorigenecity 18 gene, *ST18*, and interrupted a negative feedback loop blocking ST18 repression of its enhancer (205). Other than the aforementioned examples, no other strong evidence has been published to suggest that retrotransposons, L1 in particular, have contributed to carcinogenesis in individuals through somatic insertions. Somatic insertions may not be the only way in which retrotransposons contribute to carcinogenesis, as suggested by Shukla and colleagues in a study on hepatocellular carcinoma (205). In 4/19 patients evaluated, germline insertions into the tumor suppressor *MCC*, mutated in colon cancers, and subsequent absence of

MCC expression was confirmed (205). Clearly, it is possible for retrotransposons to contribute to cancer development; however, the question remains- how often does it occur?

When studying a specific disease such as esophageal carcinoma, we must work backwards starting with the cancerous tissue sample and attempt to reconstruct the process which lead to its existence; however, this is far from an ideal way to determine if retrotransposition contributes to cancer development. Even when insertions are detected in key genes, extensive functional studies need to be performed to show that the gene itself is being aberrantly expressed. Even if the gene is not being expressed as it should be, it is often still unclear how it's overexpression or under-expression lead to the disease developing. In fact, many of the cancer genes listed in databases such as the 'Network of Cancer Genes' are only cancer-associated genes since they have not been definitely shown to contribute to tumorigenesis (285).

It would be ideal to model what occurs following the disruption of cellular retrotransposition control and whether or not the increased activity results in cancer. A study, like the aforementioned one, could be done in an animal model, although modeling the dysregulation of the many pathways which modulate retrotransposition would be arduous and potentially cost prohibitive. It would be necessary to not only model the dysregulation of each pathway controlling retrotransposition individually, but also to model them in various combinations. Due to the complex nature of modeling the many pathways contributing to the tight retrotransposon regulation, it would be much easier to accomplish such a study in cell culture. After finding captivating evidence in a cell culture model that one or more pathways in concert promote retrotransposons to participate in carcinogenesis, the experiment could then be repeated in an animal model for confirmation. Essentially, a great deal of work remains to be

accomplished in order to plainly show that retrotransposons participate in cancer development on a semi-regular or regular basis.

Understanding individual variation and evaluating retrotransposon activity in histologically normal tissue

Recent findings from our lab and others have confirmed that insertions occur in normal GI tissues such as liver, colon, and esophagus occur and in some cases are then amplified in a tumor of the same tissue (204–206). The key to understanding retrotransposon's contribution to carcinogenesis may be in understanding the frequency at which elements mobilize in normal tissue and are subsequently selected for amplification. To acquire a better estimate of retrotransposon activity in histologically normal tissue, a significantly larger number of individuals need to be evaluated. Missing sub-clonal somatic insertions present in normal tissue could also be due to the methods used to validate predicted insertions. For example, in our work, we check every validated somatic insertion with a nested PCR in both the tumor and in the normal. We have shown that our nested PCR amplifies as little as one insertion in a thousand cells (204). Unfortunately, only our lab has utilized nested PCR to robustly determine whether or not insertions validated in tumor samples are truly absent from the matched normal tissue even at a low concentration (204,206).

To date, 598 cancer and/or precancerous condition patients' genomes have been evaluated for retrotransposon activity; however, in only 38 of these cases nested PCR was utilized to ensure somatic insertions detected were truly absent from the normal tissue (201–206,210,244). For about 10% of the patients evaluated, 61 of 598, blood was utilized for the matched normal control in the study. Using blood as the normal tissue in a study of retrotransposition is far from ideal because sub-clonal insertions in the matched normal tissue

will be missed entirely. Even using nested PCR, we are limited in the sensitivity we have to detect sub-clonal insertions as we cannot detect insertions which are at a lower concentration than 1/1000 in a sample (204).

Knowing the rate of sub-clonal insertions in the normal tissue will allow us to determine the likelihood that an insertion will occur into a gene whose up-regulation or down-regulation might confer a selective advantage upon a cell and lead to carcinogenesis. Single cell sequencing would be an ideal way to increase our sensitivity to detect insertions which are rare in the normal tissue. Because single cell sequencing may be cost prohibitive for looking at thousands of cells from each individuals studied, nested PCR could initially be used to screen individuals who have sub-clonal insertions which are detectable at 1 in 1000 cells. After individuals who are known to have sub-clonal insertions in normal tissues have been identified, their samples could be single cell sequenced to determine the frequency with which the sub-clonal insertions occurred in their normal tissue.

It has been evident from the many studies performed to date that retrotransposition is active to different degrees in different individuals. Among individuals with somatic insertions in their tumors, there are 'jackpot' winners who likely have hundred or even thousands of somatic insertions and there are individuals who do not possess a single detectable somatic insertion. The variation among individuals with regard to the activity of their mobile DNA is largely a mystery even though many pathways which restrict these elements are known. It has been shown that methylation of L1 promotors is inversely correlated with their expression in cancer patients in vivo; however, L1 expression does not necessarily mean that retrotransposition is occurring. In order to characterize what factors play a role in the successful retrotransposition of elements in cancer genomes, more individuals must be evaluated. When an individual has many

validated somatic insertions, he or she should undergo an expression profile including all pertinent genes involved in mobile DNA suppression in humans. Only by evaluating a larger number of individuals and aggregating the data can we hope to achieve an understanding of why some individuals have very active retrotransposons in their tumors and others do not.

Concluding Remarks

As reviewed in chapter 1, there is considerable prior evidence that L1 is not only a source of inter-individual genomic variation, but also active in the cancer genomes of many patients. The results presented in chapters 2 and 3 of this thesis serve as further evidence that L1 is active in cancer; furthermore, this work shows that L1 is also frequently active in normal tissues and precursor diseases to cancer as well. Although our work has contributed to the overall knowledge of L1 activity in disease, it is still uncertain to what extent it is contributing to carcinogenesis in general. In summary, these results reinforce the fact that the human genome is still susceptible to somatic insertions and the rate of acquisition of said insertions varies greatly between individuals. It will be key in the future to secure our understanding of the role of retrotransposition in all disease, not just in cancer, only then can we full appreciate its full effect on our genome.

CHAPTER 5: Appendix

Updated L1-seq protocol

Summary

L1-seq is a high-throughput sequencing technique which is utilized to identify novel L1 insertions in genomic DNA samples of interest. Using special diagnostic nucleotides unique to the youngest and most active L1 sequence, we can amplify new somatic insertions. This technique has helped to establish the number of L1 insertions present in the general population as well as the variation among individuals with regard to their complement of active L1 elements. More recently, this technique has been employed to assess the level of retrotransposition occurring in various diseases such as cancer. These efforts try to establish a connection between the process of retrotransposition and disease development and/or progression.

Introduction

Retrotransposons are nearly ubiquitous in eukaryotes from slime molds (338) to humans (111) and have contributed greatly to genome composition of these organisms. Retrotransposons make up 45% of the human genome(111). In particular, the LINE-1 (L1) element, has contributed to approximately 17% of the human genome and continues to add to it via a copy and paste mechanism with an RNA intermediate(111). L1 is the only autonomous retrotransposon in the human genome because it encodes two proteins necessary for mobilization and reinsertion into the genome; however, these two proteins, once expressed can mobilize other types of retrotransposons as well as processed psuedogenes(31,48,51). Each individual has a different complement of potentially active L1 elements; although, the majority of the L1s in each individual's genome are truncated and therefore inactive. L1-seq(249) was developed to help

characterize L1 variation among individuals because L1s have contributed to a substantial fraction of the genome and are capable of inducing many types of mutations. L1-seq has since been used to evaluate several types of cancer to establish the level of retrotransposition occurring in colon cancer, lung cancer, breast cancer, and many other cancers(201,203). Additional sequencing techniques have confirmed the L1-seq data and demonstrated that L1 elements are active in many cancer types(201,202,205,206,210,220,221). The results have demonstrated that L1s are active in a subset of patients with cancer, in addition, L1 elements are active in all epithelial cancers tested. The L1-seq technique consists of a DNA library prep as well as the validation of the predicted new insertions detected in the samples used. Although few of the insertions may be directly responsible for the development of the disease, it should be possible to utilize known insertions present in a cancer sample for monitoring the cancer's progression to metastasis. To detect metastasis using a new L1 insertion, a PCR would be performed on serum DNA from a patient to determine whether or not the insertion was detectable in the blood and therefore potentially in a floating cancer cell. This technique is useful for both evaluating the overall complement of L1 elements in a genome as well as looking for new insertion events. L1-seq utilizes unique nucleotides, 'ACA' 91-93 nucleotides from the 3' end of the element, to selectively amplify the young and active subset of elements in the human genome (6). Following the initial 5 cycles of the PCR, wherein the linear amplification of L1 elements occurs, degenerate primers are added to the mixture to exponentially amplify both polymorphic and potentially somatic insertions present in the genome.

Materials

Store all reagents as specified by manufacturers. Diligently follow all waste disposal regulations when disposing of waste materials. All primers need to be diluted to 100 µM upon receipt in Diethylpurocarbonate (DEPC) water. Primers will be further diluted as specified later in the protocol.

DNA Isolation

1. DNeasy Blood and Tissue Kit (Qiagen)

2. 500 mL of absolute ethanol (200 proof)

3. QubitTM dsDNA BR Assay kit (Life Technologies)

Library Preparation

1. Promega GoTaq Flexi

2. 25 mM MgCl2

3. 10 mM dNTPs

4. Diethylpurocarbonate (DEPC) water

5. 100% DMSO

6. Pfu polymerase

7. 1 µg of good quality DNA per sample at a concentration of 100 ng/µL (notes).

8. LE agarose

9. 1X TAE; Prepare 50X solution by dissolving 242 g Tris base in 750 mL of deionized water. Carefully add 57.1 mL glacial acetic acid, and 100 mL of 0.5 M EDTA (pH 8.0) and adjust solution to final volume of 1L. Dilute the 50X solution to 1x in deionized water.

10. Ethidium bromide (10 mg/mL)

11. QIAquick Gel Extraction Kit (Qiagen)

12. MinElute PCR Purification Kit (Qiagen)

13. Isopropanol (200 proof)

14. 500 mL of Ethanol (200 proof)

15. Agilent DNA 1000 kit or high sensitivity DNA kit (choose as needed, see notes).

16. L1-seq primers (order HPLC grade for library preparation)

Next-Generation Sequencing Data Analysis

1. Server with at least 4GB of RAM and L1-seq scripts properly formatted (https://github.com/adamewing/l1seq) 2. Bowtie2 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) and all relevant human genome files and indices as per Bowtie2 instructions.

Data Validation

1. GoTaq Green Master Mix (2X)

2. Diethylpurocarbonate DEPC water

3. DNA at concentration of 12.5 ng/ ųL

4. LE agarose

5. 1X TAE; Prepare 50X solution by dissolving 242 g Tris base in 750 mL of deionized water. Carefully add 57.1 mL glacial acetic acid, and 100 mL of 0.5 M EDTA (pH 8.0) and adjust solution to final volume of 1L. Dilute the 50X solution to 1x in deionized water.

6. Ethidium bromide (10mg/mL)

7. QIAquick Gel Extraction Kit (Qiagen)

8. Isopropanol (200 proof)

9. 500 mL of Ethanol (200 proof)

10. L1SP1A2 primer, L1nt112out, L1 "G" primer (see above)

Methods

Embedding and Cryosectioning Tissue

1. To begin, embed each piece of tissue to be assayed in OCT freezing medium. You can simply put a thin layer of the media onto a pre-chilled (< 200 C) chuck.

2. Quickly placing the thawed tissue section onto the OCT

3. Immediately cover the tissue in more OCT until it is barely visible through the medium. The OCT medium will change from clear to white, when the entire block of tissue/OCT is completely frozen, you can begin to cryosection the tissue for DNA extraction. It is best for the freezing to occur as rapidly as possible, to this end, a heat extractor can be used to enhance and shorten the freezing process and adherence to the chuck on which the tissue is being embedded in the OCT freezing medium.

4. Set the cryostat to slice sections of tissue between 10 ųm and 30 ųm.

5. During sectioning, carefully remove each roll of tissue and place 10- 20 slices into a pre-chilled (< 200 C) 1.5 mL microtube.

If the tissue sample is large enough, more than one tube of tissue slices can be made. Following the sectioning, tissue slices should be stored at -800C until it is time to isolate DNA. Embedding tissue in OCT freezing medium is only one way of extracting DNA from frozen tissue (notes).

Extracting DNA from sectioned tissue

1. Remove microtubes with tissue slices from -800 freezer and place them on ice.

2. Add 360 ųL of Buffer ATL (DNeasy Blood and Tissue kit). There is no need to further homogenize the tissue. Add 40 ųL of proteinase K and mix thoroughly by vortexing (notes).

3. Incubate at 55^0C overnight until the tissue is completely lysed. Vortex occasionally during incubation to help disperse samples.

4. Vortex for 15 seconds.

5. Add 400 µL of buffer AL (DNeasy Blood and Tissue Kit) to the sample and mix thoroughly by vortexing.

6. Immediately add 400 µL of ethanol (200 proof) and mix again thoroughly by vortexing (notes).

7. Pipette 750 µL of the mixture (including any precipitate) into the DNeasy Mini spin column placed in a 2 mL collection tube (provided in the kit).

8. Centrifuge at > 6,000 x g (8,000 rpm) for 1 minute. Discard flow through.

9. Repeat steps 7 and 8 until all of mixture has been run though the same column for each sample.

10. After the final spin with the aforementioned mixture, discard the collection tube and replace it with a new 2 mL collection tube.

11. Add 500 µL of Buffer AW1. (Ensure that ethanol has been added to Buffer AW1 before use.)

12. Centrifuge for 1 min at > 6,000 x g (8,000 rpm).

13. Discard flow-through and collection tube and place the spin column in a new 2 mL collection tube.

14. Add 500 µL of Buffer AW2. (Ensure ethanol has been added to Buffer AW2 before use.)

15. Centrifuge for 3 minutes at 20,000 x g (14,000 rpm) to dry the DNeasy membrane and then discard flow-through and collection tube.

16. Place the DNeasy Mini spin column in a clean 1.5 mL micro-centrifuge tube and pipet 100 qL of pre-warmed (550C) Buffer AE onto the DNeasy membrane.

17. Incubate at room temperature for 10 minutes.

18. Centrifuge for 1 minute at > 6,000 x g (8,000 rpm) to elute.

19. Pipette an additional 50 qL of pre-warmed AE buffer directly onto DNeasy membrane. (There is no need to replace the micro-centrifuge tube. The additional DNA elution can be collected into the same tube up to a volume of no more than 150 qL.)

20. Incubate for 10 minutes at room temperature.

21. Centrifuge for 1 minute at 20,000 x g (14,000 rpm).

This protocol is adapted from the Qiagen handbook for the DNeasy Blood and Tissue Kit (Catalog # 69504).

Measuring DNA concentration (QubitTM)

Use the QubitTM fluorometer to measure DNA concentration because it is one of the most accurate methods. Follow manufacturer protocols exactly (http://www.ebc.uu.se/digitalAssets/176/176882_3qubitquickrefcard.pdf),

L1-seq library preparation

1. Before beginning the relevant library prep PCRs, it is necessary to determine which samples will be pooled together in the library preparation. As many as ten samples can be pooled together

without using barcoding (notes). Equal amounts of each sample must be put into the DNA pool to be used in the library. A total of 24 µL of pooled DNA at a concentration of 100 ng/µL is needed. If there is not enough DNA available from one or more samples, a modified protocol can be used (notes).

2. Round 1 PCR: Linear amplification of L1 flanks followed by a hemi-specific PCR incorporating the Illumina sequencing primer (Fig 5.1). To prevent running out of master mix, make enough for 9 reactions even though only 8 reactions will be assembled. Master mix (per 1 reaction): 10 µL of Promega Go-Taq flexi buffer (5X), 6 µL of MgCl2 25 mM, 2 µL of L1SP1A2 primer 20 µM, (the second primer for the reaction, the degenerate primers previously mentioned, will be added following the completion of 5 cycles of the PCR which consists of the linear amplification step), 0.5 µL of DMSO 100%, 0.5 µL of FlexiTaq, 1 µL of dNTPs 10 mM, 2 µL of pooled DNA (at 100 ng/µL), 24 µL of DEPC water. The reaction should total 46 µL before the addition of 4 µL of the degenerate primers to be added after linear amplification is finished. When the linear amplification is finished, add 1 of each of the degenerate primers at 5 µM (e.g. DEGSeq1N5TCTGT) to each of the 8 reactions. Use the following cycling program:

L1-Seq PCR 1:

1) 95 C for 2 minutes 30 seconds

2) 95 C for 30 seconds

3) 58 C for 1 minute

4) 72 C for 2 minutes

5) Go to step 2 (5x)

6) 60 C (pause and add 4 µL of degenerate primer into each of the 8 reactions, one primer per reaction for each of the 8 different degenerate primers)

7) 95 C for 30 seconds

8) 55 C for 30 seconds

9) 72 C for 1 minute and 30 seconds

10) Go to step 7 (14x)

11) 72 C for 10 minutes

12) 4 C hold

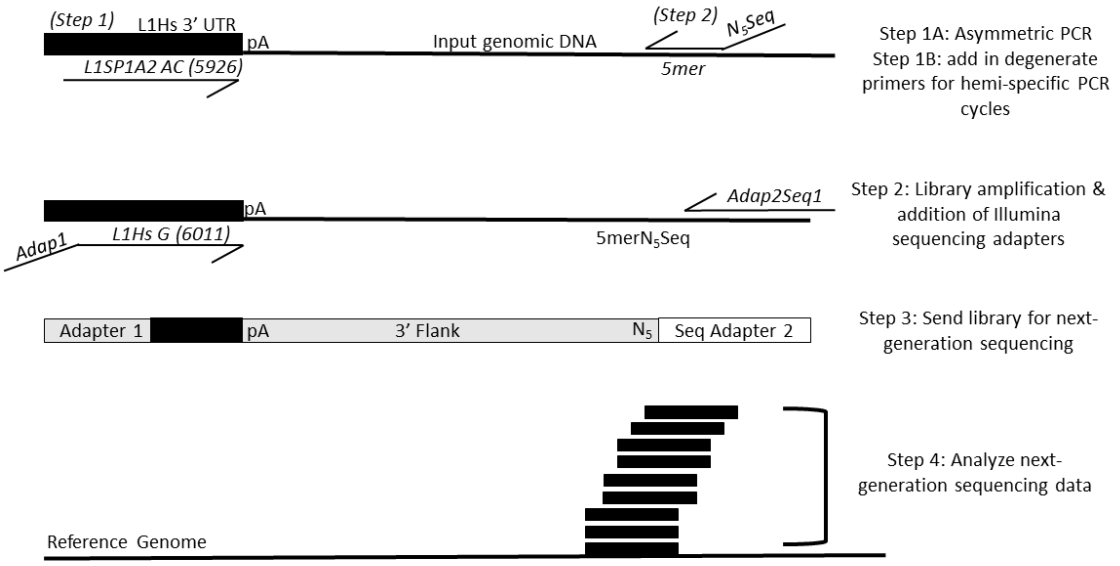


Figure 5.1: The first PCR consists of five cycles which enrich for sequences containing human-specific L1 sequences via primer extension with the L1Sp1A2 primer using diagnostic nucleoties for the human specific subfamily of L1. After enrichment for human-specific L1 flanks in the first 5 cycles, each reaction has one degenerate primer added. There are 8 degenerate primers, each with a specified 5-mer at the 3′ end preceded by five degenerate bases (NNNNN) and a sequencing primer used for the Illumina platform. Eight different reactions are performed, each with a unique 5mer. The second PCR enriches for human-specific L1 3′ flanks by utilizing another diagnostic nucleotide in the 3' UTR of L1 and adds the necessary adapter sequences via primer overhangs. The products of this reaction are mixed in equimolar ratios before sequencing on the Illumina 2500 platform. Following sequencing and initial processing, the reads are aligned to the human genome (hg19) and L1 insertions are identified for validation.

3. Purify all 8 reactions on 8 separate Qiagen PCR Clean-up columns following the Qiagen protocol, eluting in 50 µL of pre-warmed (55 C) EB (do a 10-minute final incubation before elution to optimize DNA eluted from column).

4. L1-Seq PCR 2: Amplification of library and addition of the Illumina sequencing adapters (Fig. 1). Master mix (for 1 reaction): 12.5 µL of master mix (Promega GoTaq Green 2X), 1.5 µL of primer Adap1L1HsG 20 mM, 1.5 µL of Adap2Seq1 20 µM, 2.5 µL of purified round 1 product (1 degenerate primer per reaction), 7 µL of DEPC ddH2O. Again, make enough for 9 reactions to prevent running out of master mix for the samples to be amplified. The reactions will each have a total of 25 µL. Use the following cycling program:

L1-seq PCR 2

1) 95 C for 2 minutes

2) 95 C for 30 seconds

3) 62 C for 30 seconds

4) 72 C for 1 minute

5) Go to 2 (19x)

6) 72 C for 5 minutes

7) 4 C Hold

5. Resolve products on a 1% TAE gel.

6. Excise the constellation of bands between 200- 500 nucleotides with a sterile scalpel (using a different scalpel for each reaction) and purify the DNA using the Qiagen Gel Purification protocol.

7. Elute the library in 50 qL of pre-warmed EB buffer (55C with a 10-minute incubation before elution).

8. Run each DNA sample on the Agilent Bioanalyzer with the DNA 1000 kit to get an accurate measure of concentration and the average size of the DNA amplified. Using the concentration and average size of the molecules, calculate how to add the DNA from all 8 reactions in equimolar ratios to one tube.

9. After mixing the DNAs together, purify the entire mixture with the Qiagen MinElute PCR purification kit eluting in 50 qL of pre-warmed EB (55C and 10-minute incubation at room temperature before elution).

10. End-polishing must be performed on the library because Taq leaves adenine overhangs which could cause problems when the library is annealed to the Illumina flow cell. To accomplish the end-polishing: mix 6 qL of 10x Pfu buffer, 2.5 qL of Pfu polymerase, 2.5 qL of dNTPs 10 mM, and 49 qL of library. Incubate for one hour at 72C.

11. Purify reaction on a Qiagen MinElute column and elute in 10 qL of pre-warmed (55C) EB following a 10-minute room-temperature incubation before elution.

12. Measure final DNA concentration with QubitTM fluorometer to get an accurate concentration for sending samples for next-generation sequencing on the Illumina HiSeq 2500. Opt for single end sequencing and at least 100 bp reads.

3.5 Analyzing the Next-generation sequencing data

1. Obtain the sequencing reads from the core or center where the samples were sequenced and transfer them into the L1-seq directory on the server in which all the correctly formatted L1-seq scripts reside. These scripts can be acquired from https://github.com/adamewing/l1seq.

2. Obtain the contents of a database with all reference L1 insertions and polymorphic L1 insertions which have been previously published to use for filtering sequencing data. A database of L1 insertions may be obtained at: http://nar.oxfordjournals.org/content/43/D1/D43(339).

3. Once the documents are downloaded through the terminal, they must be unzipped. To unzip the fastq.gz files type "gunzip –d FILE_NAME.fastq.gz &" (the & symbol allows the unzipping to run in the background so that you can set all the files and can unzip simultaneously by typing this command for each file in turn.)

4. After the files are unzipped, use the script to run bowtie and create indices for your data. You can execute this process with the command "./run_bowtie.py/whatever_fastq /directions/to/bowtie directions/to/hg19.fa &". For any process which takes more than 10 minutes, it is helpful to use the screen function by typing "screen –rAad" which will allow for the monitoring of all the processes simultaneously running. It also enables the user to monitor the total memory being used for analysis and the length of time each process has been running.

5. Once the run_bowtie script finishes, run the l1seq.py script as follows: "./l1seq.py –bam whatever.ba, > whatever.l1seq.txt &"

6. Once all of the L1-seq.txt files are made, all the files need to be compressed for sorting. To compress the files, type "bgzip whatever_l1seq.txt &"

7. To sort the files, type "tabix –s 1 –b 3 –e 4 watever.l1seq.txt.gz &"

8. All the files must be compared to one another (e.g. normal compared with tumor etc.) to do this analysis, type "./compare.py group1_L1seq.txt.gz group2_l1seq.txt.gz group3_l1seq.txt.gz > filename_for_comparisons.tsv &"

9. Finally, after the comparison file has been made, primers must be designed for validation of the data. It is best to run the makeprimers.pl script on the entire comparison file before looking at the data because the script does not take long to run and having the primer sequences ready to order is very useful. To run this final script, type "./makeprimers.pl filename_for_comparisons.tsv > filename_for_comparisons_with_primers.tsv &". Use sftp to transfer the files back to your local computer if desired.

Validating the Predicted Insertions from L1-seq with Site Specific PCR

1. The presence of nonreference insertions is validated with site specific PCR (Figure 5.1). If the samples are not barcoded, all samples in a pool must be evaluated for the presence or absence of a predicted insertion. The DNA from each input sample from a pool needs to be at 12.5 ng/ uL and 2 uL of DNA used per reaction. The primers will be named by the makeprimers.pl script as "filled site" or FS and "empty site" ES refer to figure 5.1 for orientation of primers with regard

to the potential insertion. For each validation to be complete, the FS and L1SP1A2 primer reaction needs to be performed on all samples in the pool from which the prediction came. If comparing two states of the same tissue such as tumor and normal and the insertion is predicted only in one, the reaction must also be performed on both DNA samples. A control reaction can also be performed with the "empty site" or ES primer and the FS primer. Both the FS and ES primers are genomic and will produce a product of predetermined size in any DNA sample regardless of presence or absence of an insertion.

2. For the FS/L1SP1A2 (filled site PCR) use the following master mix (1x): 12.5 µL of Promega GoTaq green (2X) master mix, 0.8 µL of FS primer 20 µM, 1.6 µL of L1SP1A2 20 µM, 2 µL of genomic DNA (12.5 ng/µL), 8.2 µL of DNase free H2O. For the FS/ES primers (empty site PCR) use the following master mix (1X): 12.5 µL of Promega GoTaq green (2X) master mix, 1 µL of FS primer 20 µM, 1 µL of ES primer 20 µM, 2 µL of genomic DNA (12.5 ng/µL), 9.5 µL of DNase free H2O. Use the following parameters for the PCR:

3' L1 Validation PCR

1) 95 C for 2 minutes

2) 95 C for 30 seconds

3) 57 C for 30 seconds

4) 72 C for 1 minute and 30 seconds

5) Go to step 2 (29x)

6) 72 C for 5 minutes

7) 4 C Hold

3. Run the PCR products on a 1.5% TAE gel to resolve the products. Take images of the gel while it is exposed to UV to visualize the products. Excise fragments which are unique to only one of the samples upon which the PCR was run (Fig 5.2B). Isolate the DNA from the band and send for sequencing. If no clear filled site band is uniquely present in one of the samples tested, a nested PCR may be necessary (Fig 5.2A). Alternatively, the PCR conditions can be further optimized to attempt to amplify the insertion.

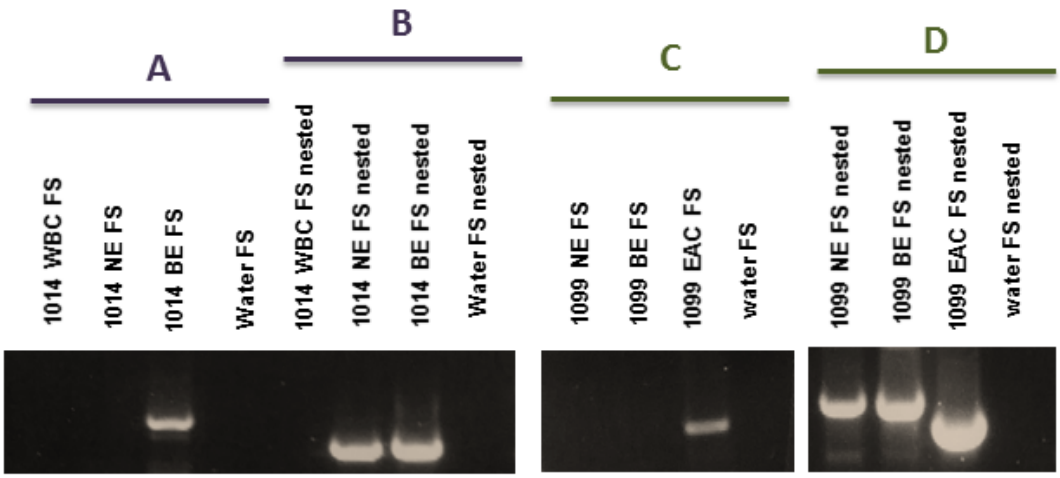


Figure 5.2: A) Diagram of the PCR validation scheme for putative insertions, the 3' end of the LINE-1 insertion is pictured adjacent to a poly A tail. The nested empty site and filled site primers are flanking the empty and filled site primers. In a nested PCR, the nested primers are used in the first of two reactions. One and a half uL of product from the first reaction (with ES and FS primers) are used as template in a second PCR with the nested primers to amplify difficult or rare products. B) Two examples of validations for insertions present in only tumor and absent from normal DNA. On the left, a PCR result depicting both the empty site (ES) and filled site (FS) products for both the normal and tumor DNA samples from patients. Only in the tumor of patient 11 is a filled site band present confirming the insertion is present. In the image on the right side on Figure 5,2 B, a PCR depicting another validation of a somatic insertion present in BE and absent from normal esophageal and white blood cell DNA. There is only a band present in the BE sample for the FS PCR; however, the ES PCR has bands for all three DNA samples as a positive control. C) An insertion sequence with the unique genomic DNA (blue), target site duplications (purple), LINE-1 sequence (red), and the poly-A tail sequence (orange).

4. Finally, send the DNA for Sanger sequencing to ensure it is the correct product. Sequence the product with both the FS primer as well as the L1SP1A2 primer. When the sequence from the FS is aligned to the genome with BLAT or another alignment algorithm, part of the sequence should align to the genome and a poly T tract should also be visible adjacent to the aligning sequence. For the sequence from the reaction performed with the L1 specific primer, the 3' end of the L1 should be visible in addition to the poly-A tail (Fig 2C).

5. To find the 5' end of the insertion, several different methods can be utilized. Because many new L1 insertions are truncated on the 5' end of the element, it is frequently possible to detect the 5' end of the element by using the reverse complement of the L1SP1A2 primer (L1 GTG primer) with the ES primer. To do this, make the master mix as follows: mix (1x): 12.5 µL of Promega GoTaq green (2X) master mix, 0.8 µL of ES primer 20 µM, 1.6 µL of L1 GTG primer 20 µM, 2 µL of genomic DNA (12.5 ng/µL), 8.2 µL of DNase free H2O. For insertions with a longer 5' end present, this PCR will likely fail; however, it is possible to tile across the L1 element with various primers at different locations (e.g. L1nt112out) in the element accompanied by the empty site primer to find the 5' end. For this PCR, use the following master mix (1X): 12.5 µL of Promega GoTaq green (2X) master mix, 0.8 µL of ES primer 20 µM, 1.6 µL of L1 internal primer 20 µM, 2 µL of genomic DNA (12.5 ng/µL), 8.2 µL of DNase free H2O.

5' L1 GTG PCR parameters:

1) 95 C for 2 minutes

2) 95 C for 30 seconds

3) 57.5 C for 1 minute and 30 seconds

4) 72 C for 3 minutes

5) Go to Step 2 (29x)

6) 72 C for 5 minutes

7) 4 C Hold

5’ L1 internal primer (e.g. L1nt112out) PCR parameters:

1) 95 C for 2 minutes

2) 95 C for 30 seconds

3) 57 C for 30 seconds

4) 72 C for 45 seconds

5) Go to step 2 (29x)

6) 72 C for 5 minutes

7) 4 C Hold

Notes

1. If very little DNA is available for both library prep and validation PCRs, L1-seq can still be successfully performed. L1-seq has successfully been executed with as little as 25 ng of input per sample for the library prep. For the steps following next-generation sequencing, whole genome amplification can be used (e.g. the Qiagen Repli-G kit) to provide more DNA to use for the

validation PCRs. If adjusting the amount of DNA used, be sure to account for volume changes and the concentrations of the other reagents to ensure all final concentrations are the same as described in the original technique.

2. Occasionally, one of the degenerate primer reactions will not be as robust as the other reactions and when the libraries are run on a gel, the amount of DNA present is variable between reactions. This may not be an issue if there is enough DNA present after the gel purification for the samples to easily be mixed in equal amounts. However, if the concentration of the DNA isolated from the gel purification step is too little to continue without grossly diminishing the amount of total DNA in the combined library, simply repeat the second reaction of L1-seq and combine the isolated DNA from both gel purifications and concentrate the DNA. If the DNAs from the respective degenerate primer reactions are run on the Bioanalyzer and produce very different size distributions of products, it may be necessary to repeat the second L1 PCR again on that DNA sample as well. Ideally, the average product size for each degenerate primer reaction should be within one standard deviation of 350 nucleotides. If the size varies more than one standard deviation from 350, the reaction should be repeated and rerun on a gel. If the size is wrong, it is likely that the excision was initially imprecise.

3. When first performing L1-seq it is prudent to execute a TA cloning step after completing the libraries and mixing them in equimolar ratios, but before completing the end-polishing step. To do this, simply take one µL from the mixed libraries and use it in a Topo TA cloning reaction. Follow kit instructions and after growing colonies overnight, select 12 or more from each plate for colony PCR. Following colony PCR, run the product on a gel to be sure the cloning worked effectively, select some or all of the successful clones for Sanger sequencing. When analyzing

the Sanger sequencing, look for different L1Ta elements from many different areas of the genome. Essentially, this is a step to check that the library does not consist of amplicons of only a handful of LINE-1 elements in the genome and that elements in the genome are equally represented in the library. This step does not need to be performed for every library prep; however, if a problem occurs with next-generation sequencing, this step could consequently be taken to determine whether or not overrepresentation of a few elements precluded successful sequencing.

4. If validation PCRs are unsuccessful after many attempts, be sure to check the specificity of the primers being used in the amplification. Oftentimes, it is helpful to perform a nested PCR following the first conventional PCR to amplify difficult or low-copy insertions which may have been easily detectable with next-generation sequencing and not with Sanger sequencing. You can nest both the filled site primers as well as the L1Ta specific primers to increase the specificity of the reaction greatly. Nested PCR along with an increase in cycle numbers and/or altering the melting temperature of the PCR often alleviates validation PCR issues.

5. If the DNA being measured at any point in the library prep is at a low concentration and undetectable with the standard QubitTM broad range kit or the Agilent 100 DNA chip, there are low concentration versions of these reagents available.

6. Barcoding may also be utilized with this technique; however, results may vary. In 2012, Evrony et al. performed L1-seq using barcoding and were able to validate some new LINE-1 insertions following sequencing analysis. However, other groups have had more difficulty getting the technique to work well and seem to have more success with pooling samples without

barcodes. Pooling samples without barcodes does create more work for the validation steps of the technique; however, it seems to have more reproducible results.

7. With regard to choosing predicted insertions for validation, one of two main methods may be employed. A random number generator can be used to select putative somatic insertions for validation which will potentially give a good estimate of the number of true somatic insertions in the dataset. Alternatively, putative somatic insertions with unique read counts above 5, map scores of 1, and alignment windows of at least 100 base pairs can be selected for validation. Depending on the validation rate with the primary insertions selected, the level of stringency can be altered until the ideal validation rate is achieved. A validation rate above 60% is generally acceptable for this technique; however, PCR optimization, good primer design, and good DNA are key to successful validations.

Bibliography

1. Britten RJ, Kohne DE. Repeated Sequences in DNA. Science. 1968. p. 529–40.

2. McClintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36(6):344–55.

3. McClintock B. Induction of Instability at Selected Loci in Maize. Genetics. 1953;38(6):579–99.

4. Lander E, Linton L, Birren B, Nusbaum C, Zody M, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.

5. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304–51.

6. Kleckner N, Chan RK, Tye BK BD. Mutagenesis by insertion of a drug resistance element carrying an inverted repetition. J Mol Biol. 1975;97(4):561–75.

7. Brügger K, Redder P, She Q, Confalonieri F, Zivanovic Y, Garrett RA. Mobile elements in archaeal genomes. FEMS Microbiology Letters. 2002. p. 131–41.

8. Boyd EF, Hartl DL. Nonrandom location of IS1 elements in the genomes of natural isolates of Escherichia coli. Mol Biol Evol. 1997;14(7):725–32.

9. Craig NL. Unity in transposition reactions. Science. 1995;270(5234):253–4.

10. Craig NL. Target site selection in transposition. Annu Rev Biochem. 1997;66:437–74.

11. Kapitonov V V, Jurka J. Rolling-circle transposons in eukaryotes. Proc Natl Acad Sci U S A. 2001;98(15):8714–9.

12. Kapitonov V V, Jurka J. Self-synthesizing DNA transposons in eukaryotes. Proc Natl Acad Sci U S A. 2006;103(12):4540–5.

13. Pritham EJ, Putliwala T, Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene. 2007;390(1-2):3–17.

14. Boeke JD, Garfinkel DJ, Styles CA, Fink GR. Ty elements transpose through an RNA intermediate. Cell. 1985;40(3):491–500.

15. Finnegan DJ, Rubin GM, Young MW, Hogness DS. Repeated gene families in Drosophila melanogaster. Cold Spring Harb Symp Quant Biol. 1977;42(2):1053–63.

16. Lueders KK, Kuff EL. Sequences associated with intracisternal A particles are reiterated in the mouse genome. Cell. 1977;12(4):963–72.

17. Mills RE, Bennett EA, Iskow RC, Devine SE. Which transposable elements are active in the human genome? Trends in Genetics. 2007. p. 183–91.

18. Curcio MJ BM. Retrohoming: cDNA mediated mobility of group II introns requires a catalytic RNA. Cell. 1996;84(1):9–12.

19. Beauregard A, Curcio MJ, Belfort M. The take and give between retrotransposable elements and their hosts. Annu Rev Genet. 2008;42:587–617.

20. Pardue M-L, DeBaryshe PG. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu Rev Genet. 2003;37:485–511.

21. Scott AF, Schmeckpeper BJ, Abdelrazik M, Comey CT, O'Hara B, Rossiter JP, et al. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics. 1987;1(2):113–25.

22. Skowronski J, Singer MF. Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proc Natl Acad Sci U S A. 1985;82(18):6050–4.

23. Swergold GD. Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol. 1990;10(12):6718–29.

24. Martin SL, Bushman FD. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. Mol Cell Biol. 2001;21(2):467–75.

25. Moran J V, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH. High frequency retrotransposition in cultured mammalian cells. Cell [Internet]. 1996 Nov 29;87(5):917–27. Available from: http://www.sciencedirect.com/science/article/pii/S0092867400819984

26. Kulpa DA, Moran J V. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. Hum Mol Genet. 2005;14(21):3237–48.

27. Martin SL, Cruceanu M, Branciforte D, Li PWL, Kwok SC, Hodges RS, et al. LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 Protein. J Mol Biol. 2005;348(3):549–61.

28. Martin SL, Bushman D, Wang F, Li PWL, Walker A, Cummiskey J, et al. A single amino acid substitution in ORF1 dramatically decreases L1 retrotransposition and provides insight into nucleic acid chaperone activity. Nucleic Acids Res. 2008;36(18):5845–54.

29. Martin SL. Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. RNA Biol. 2010;7(6):706–11.

30. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, et al. Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol. 2001;21(4):1429–39.

31. Esnault C, Maestre J, Heidmann T. Human LINE retrotransposons generate processed

pseudogenes. Nat Genet. 2000;24(4):363–7.

32. Khazina E, Truffault V, Büttner R, Schmidt S, Coles M, Weichenrieder O. Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. Nat Struct Mol Biol. 2011;18(9):1006–14.

33. Dombroski BA, Feng Q, Mathias SL, Sassaman DM, Scott AF, Kazazian HH, et al. An in vivo assay for the reverse transcriptase of human retrotransposon L1 in Saccharomyces cerevisiae. Mol Cell Biol. 1994;14(7):4485–92.

34. Feng Q, Moran J V., Kazazian HH, Boeke JD. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell. 1996;87(5):905–16.

35. Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, et al. The genomic complexity of primary human prostate cancer. Nature. 2011;470(7333):214–20.

36. Mathias SL, Scott AF, Kazazian HH, Boeke JD, Gabriel A. Reverse transcriptase encoded by a human transposable element. Science. 1991;254(5039):1808–10.

37. Xiong Y, Eickbush TH. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. 1990;9(10):3353–62.

38. Fanning T, Singer M. to retrovirus proteins Interspersed repetitive DNA sequences make up a substantial portion of Nucleic Acids Research. 1987;15(5):2251–60.

39. Clements AP, Singer MF. The human LINE-1 reverse transcriptase: Effect of deletions outside the common reverse transcriptase domain. Nucleic Acids Res. 1998;26(15):3528–35.

40. Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran J V. Unconventional translation

of mammalian LINE-1 retrotransposons. Genes Dev. 2006;20(2):210–24.

41. Li PWL, Li J, Timmerman SL, Krushel LA, Martin SL. The dicistronic RNA from the mouse LINE-1 retrotransposon contains an internal ribosome entry site upstream of each ORF: Implications for retrotransposition. Nucleic Acids Res. 2006;34(3):853–64.

42. Babushok D V., Kazazian HH. Progress in understanding the biology of the human mutagen LINE-1. Human Mutation. 2007. p. 527–39.

43. Luan DD, Korman MH, Jakubczak JL ET. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell. 1993;72(4):595–605.

44. Gasior SL, Wakeman TP, Xu B, Deininger PL. The human LINE-1 retrotransposon creates DNA double-strand breaks. J Mol Biol. 2006;357(5):1383–93.

45. Ostertag EM, Kazazian H.H. J. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. 2001;11(12):2059–65.

46. Callinan PA, Wang J, Herke SW, Garber RK, Liang P, Batzer MA. Alu retrotransposition-mediated deletion. J Mol Biol. 2005;348(4):791–800.

47. Kriegs JO, Churakov G, Jurka J, Brosius J, Schmitz J. Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. Trends in Genetics. 2007. p. 158–61.

48. Dewannieux M, Esnault C, Heidmann T. LINE-mediated retrotransposition of marked Alu sequences. Nat Genet. 2003;35(1):41–8.

49. Ostertag EM, Kazazian HH. Biology of mammalian L1 retrotransposons. Annu Rev Genet. 2001;35:501–38.

50. Wang H, Xing J, Grover D, Hedges Kyudong Han DJ, Walker JA, Batzer MA. SVA

elements: A hominid-specific retroposon family. J Mol Biol. 2005;354(4):994–1007.

51. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH. SVA elements are nonautonomous retrotransposons that cause disease in humans. Am J Hum Genet. 2003;73(6):1444–51.

52. Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA. Emergence of primate genes by retrotransposon-mediated sequence transduction. Proc Natl Acad Sci U S A. 2006;103(47):17608–13.

53. Hancks DC, Ewing AD, Chen JE, Tokunaga K, Kazazian HH. Exon-trapping mediated by the human retrotransposon SVA. Genome Res. 2009;19(11):1983–91.

54. Damert A, Raiz J, Horn A V., Löwer J, Wang H, Xing J, et al. 5́-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. Genome Res. 2009;19(11):1992–2008.

55. Hancks DC, Mandal PK, Cheung LE, Kazazian HH. The Minimal Active Human SVA Retrotransposon Requires Only the 5'-Hexamer and Alu-Like Domains. Molecular and Cellular Biology. 2012. p. 4718–26.

56. Smit AF, Tóth G, Riggs AD, Jurka J. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol. 1995;246(3):401–17.

57. Khan H, Smit A, Boissinot S. Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. 2006;16(1):78–87.

58. Boissinot S, Furano A V. Adaptive evolution in LINE-1 retrotransposons. Mol Biol Evol. 2001;18(12):2186–94.

59. L VV. A new evolutionary theory. Evol Theory. 1973;1(1):1–30.

60. Chen H, Lilley CE, Yu Q, Lee D V., Chou J, Narvaiza I, et al. APOBEC3A is a potent

inhibitor of adeno-associated virus and retrotransposons. Curr Biol. 2006;16(5):480–5.

61. Bogerd HP, Wiegand HL, Hulme AE, Garcia-Perez JL, O'Shea KS, Moran J V, et al. Cellular inhibitors of long interspersed element 1 and Alu retrotransposition. Proc Natl Acad Sci U S A. 2006;103(23):8780–5.

62. Muckenfuss H, Hamdorf M, Held U, Perkovic M, Löwer J, Cichutek K, et al. APOBEC3 proteins inhibit human LINE-1 retrotransposition. J Biol Chem. 2006;281(31):22161–72.

63. Stenglein MD, Harris RS. APOBEC3B and APOBEC3F inhibit L1 retrotransposition by a DNA deamination-independent mechanism. J Biol Chem. 2006;281(25):16837–41.

64. Bourc'his D, Bestor TH. Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. Nature. 2004;431(7004):96–9.

65. Kato Y, Kaneda M, Hata K, Kumaki K, Hisano M, Kohara Y, et al. Role of the Dnmt3 family in de novo methylation of imprinted and repetitive sequences during male germ cell development in the mouse. Hum Mol Genet. 2007;16(19):2272–80.

66. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, et al. A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. Mol Cell. 2008;31(6):785–99.

67. Woodcock DM, Lawler CB, Linsenmeyer ME, Doherty JP, Warren WD. Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. J Biol Chem. 1997;272(12):7810–6.

68. Yoder JA, Walsh CP, Bestor TH. Cytosine methylation and the ecology of intragenomic parasites. Trends in Genetics. 1997. p. 335–40.

69. Walsh CP, Chaillet JR, Bestor TH. Transcription of IAP endogenous retroviruses is

constrained by cytosine methylation. Nat Genet. 1998;20(2):116–7.

70. Baba Y, Huttenhower C, Nosho K, Tanaka N, Shima K, Hazra A, et al. Epigenomic diversity of colorectal cancer indicated by LINE-1 methylation in a database of 869 tumors. Mol Cancer. 2010;9:125.

71. Muotri AR, Marchetto MCN, Coufal NG, Oefner R, Yeo G, Nakashima K, et al. L1 retrotransposition in neurons is modulated by MeCP2. Nature. 2010;468(7322):443–6.

72. Perng W, Mora-Plazas M, Marín C, Rozek LS, Baylin A, Villamor E. A Prospective Study of LINE-1DNA Methylation and Development of Adiposity in School-Age Children. PLoS One. 2013;8(4):1–7.

73. Aravin A, Gaidatzis D, Pfeffer S, Lagos-Quintana M, Landgraf P, Iovino N, et al. A novel class of small RNAs bind to MILI protein in mouse testes. Nature. 2006;442(7099):203–7.

74. Heras SR, Macias S, Plass M, Fernandez N, Cano D, Eyras E, et al. The Microprocessor controls the activity of mammalian retrotransposons. Nat Struct Mol Biol [Internet]. 2013;20(10):1173–81. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3836241&tool=pmcentrez&rendertype=abstract

75. Macias S, Plass M, Stajuda A, Michlewski G, Eyras E, Cáceres JF. DGCR8 HITS-CLIP reveals novel functions for the Microprocessor. Nature Structural & Molecular Biology. 2012. p. 760–6.

76. Sijen T, Plasterk RHA. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. Nature. 2003;426(6964):310–4.

77. Watanabe T, Takeda A, Tsukiyama T, Mise K, Okuno T, Sasaki H, et al. Identification

and characterization of two novel classes of small RNAs in the mouse germline: Retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. Genes Dev. 2006;20(13):1732–43.

78. Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, et al. LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell. 2010;141(6):956–69.

79. Levin HL, Moran J V. Dynamic interactions between transposable elements and their hosts. Nat Rev Genet. 2011;12(9):615–27.

80. Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature. 2008;453(7194):539–43.

81. Yang N, Kazazian HH. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat Struct Mol Biol. 2006;13(9):763–71.

82. Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, et al. An endogenous small interfering RNA pathway in Drosophila. Nature. 2008;453(7196):798–802.

83. Malone CD, Hannon GJ. Small RNAs as Guardians of the Genome. Cell. 2009. p. 656–68.

84. Slotkin RK, Vaughn M, Borges F, Tanurdzić M, Becker JD, Feijó JA, et al. Epigenetic reprogramming and small RNA silencing of transposable elements in pollen. Cell. 2009;136(3):461–72.

85. Arjan-Odedra S, Swanson CM, Sherer NM, Wolinsky SM, Malim MH. Endogenous MOV10 inhibits the retrotransposition of endogenous retroelements but not the replication

of exogenous retroviruses. Retrovirology. 2012. p. 53.

86. Goodier JL, Cheung LE, Kazazian HH. MOV10 RNA Helicase Is a Potent Inhibitor of Retrotransposition in Cells. PLoS Genet. 2012;8(10).

87. Goodier JL, Cheung LE, Kazazian HH. Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition. Nucleic Acids Res. 2013;41(15):7401–19.

88. Liu C, Zhang X, Huang F, Yang B, Li J, Liu B, et al. APOBEC3G inhibits microRNA-mediated repression of translation by interfering with the interaction between Argonaute-2 and MOV10. J Biol Chem. 2012;287(35):29373–83.

89. Stetson DB, Ko JS, Heidmann T, Medzhitov R. Trex1 Prevents Cell-Intrinsic Initiation of Autoimmunity. Cell. 2008;134(4):587–98.

90. Bogerd HP, Wiegand HL, Doehle BP, Lueders KK, Cullen BR. APOBEC3A and APOBEC3B are potent inhibitors of LTR-retrotransposon function in human cells. Nucleic Acids Res. 2006;34(1):89–95.

91. Casavant NC, Hardies SC. The dynamics of murine LINE-1 subfamily amplification. J Mol Biol. 1994;241(3):390–7.

92. Cabot EL, Angeletti B, Usdin K, Furano A V. Rapid evolution of a young L1 (LINE-1) clade in recently speciated Rattus taxa. J Mol Evol. 1997;45(4):412–23.

93. Goodier JL, Ostertag EM, Du K, Kazazian H.H. J. A novel active L1 retrotransposon subfamily in the mouse. Genome Res. 2001;11(10):1677–85.

94. Skowronski J, Fanning TG, Singer MF. Unit-length line-1 transcripts in human teratocarcinoma cells. Mol Cell Biol. 1988;8(4):1385–97.

95. Boissinot S, Chevret P, Furano A V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. Mol Biol Evol. 2000;17(6):915–28.

96. Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, et al. A comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet. 2002;71(2):312–26.

97. Grimaldi G, Skowronski J, Singer MF. Defining the beginning and end of KpnI family segments. EMBO J. 1984;3(8):1753–9.

98. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran J V, et al. Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A. 2003;100(9):5280–5.

99. Mills RE, Bennett EA, Iskow RC, Devine SE. Which transposable elements are active in the human genome? Trends Genet. 2007;23(4):183–91.

100. Bennett EA, Keller H, Mills RE, Schmidt S, Moran J V., Weichenrieder O, et al. Active Alu retrotransposons in the human genome. Genome Res. 2008;18(12):1875–83.

101. Seleme M del C, Vetter MR, Cordaux R, Bastone L, Batzer MA, Kazazian HH. Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc Natl Acad Sci U S A. 2006;103(17):6611–6.

102. Lutz SM, Vincent BJ, Kazazian HH, Batzer MA, Moran J V. Allelic heterogeneity in LINE-1 retrotransposition activity. Am J Hum Genet. 2003;73(6):1431–7.

103. Cheung VG, Cheung VG, Spielman RS, Spielman RS, Ewens KG, Ewens KG, et al. Mapping determinants of human gene expression by regional and genome-wide association. Nature. 2005;437(7063):1365–9.

104. Boissinot S, Entezam A, Young L, Munson PJ, Furano A V. The insertional history of an active family of L1 retrotransposons in humans. Genome Res. 2004;14(7):1221–31.

105. Sheen F, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, et al. Reading between the LINEs: Human Genomic Variation Induced by LINE-1 Retrotransposition. Genome Res. 2000;10(10):1496–508.

106. Batzer M a, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002;3(5):370–9.

107. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. dbRIP: A highly integrated database of retrotransposon insertion polymorphisms in humans. Human Mutation. 2006. p. 323–9.

108. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, et al. LINE-1 retrotransposition activity in human genomes. Cell. 2010;141(7):1159–70.

109. Kazazian HH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. Nature. 1988;332(6160):164–6.

110. Beck CR, Garcia-Perez JL, Badge RM, Moran J V. LINE-1 Elements in Structural Variation and Disease. Annu Rev Genomics Hum Genet. 2011;12:187–215.

111. Hancks DC, Kazazian HH. Active human retrotransposons: Variation and disease. Current Opinion in Genetics and Development. 2012. p. 191–203.

112. Kaer K, Speek M. Retroelements in human disease. Gene. 2013. p. 231–41.

113. Kutsche K, Ressler B, Katzera HG, Orth U, Gillessen-Kaesbach G, Morlot S, et al. Characterization of breakpoint sequences of five rearrangements in L1CAM and ABCD1

(ALD) genes. Hum Mutat. 2002;19(5):526–35.

114. Gu Y, Kodama H, Watanabe S, Kikuchi N, Ishitsuka I, Ozawa H, et al. The first reported case of Menkes disease caused by an Alu insertion mutation. Brain Dev [Internet]. 2007;29(2):105–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17178205

115. Conley ME, Partain JD, Norland SM, Shurtleff S a, Kazazian HH. Two independent retrotransposon insertions at the same site within the coding region of BTK. Hum Mutat. 2005;25(3):324–5.

116. Apoil P a, Kuhlein E, Robert a, Rubie H, Blancher a. HIGM syndrome caused by insertion of an AluYb8 element in exon 1 of the CD40LG gene. Immunogenetics [Internet]. 2007;59(1):17–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17146684

117. Claverie-Martin F, González-Acosta H, Flores C, Antón-Gamero M, García-Nieto V. De novo insertion of an Alu sequence in the coding region of the CLCN5 gene results in Dent's disease. Hum Genet [Internet]. 2003;113(6):480–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14569459

118. Masson E, Hammel P, Garceau C, Bénech C, Quéméner-Redon S, Chen J-M, et al. Characterization of two deletions of the CTRC locus. Mol Genet Metab [Internet]. 2013;109(3):296–300. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23721890

119. Sukarova E, Dimovski AJ, Tchacarova P, Petkov GH, Efremov GD. An Alu Insert as the Cause of a Severe Form of Hemophilia A. Acta Haematol. 2001;106:126–9.

120. Ganguly A, Dunbar T, Chen P, Godmilow L, Ganguly T. Exon skipping caused by an intronic insertion of a young Alu Yb9 element leads to severe hemophilia A. Hum Genet

[Internet]. 2003;113(4):348–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12884004

121. Green PM, Bagnall RD, Waseem NH, Giannelli F. Haemophilia A mutations in the UK: results of screening one-third of the population. Br J Haematol [Internet]. 2008;143(1):115–28. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18691168

122. Vidaud D, Tartary M, Costa J. substitutions at the-6 position in the promoter region of the factor IX gene result in different severity of hemophilia B Leyden: consequences for genetic counseling. Hum Genet [Internet]. 1993;241–4. Available from: http://link.springer.com/article/10.1007/BF00218264

123. Wulff K, Gazda H, Schröder W, Robicka-Milewska R, Herrmann FH. Identification of a novel large F9 gene mutation-an insertion of an Alu repeated DNA element in exon e of the factor 9 gene. Hum Mutat [Internet]. 2000 Mar;15(3):299. Available from: http://onlinelibrary.wiley.com/store/10.1002/humu.49/asset/49_ftp.pdf;jsessionid=D5C796D124830106EF1D308CFE5EA8F8.f04t03?v=1&t=hvv8j4da&s=453903ac92ca83cd3410ec342a11df670ef6d15c

124. Li X, Scaringe W a, Hill K a, Roberts S, Mengos a, Careri D, et al. Frequency of recent retrotransposition events in the human factor IX gene. Hum Mutat [Internet]. 2001;17(6):511–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11385709

125. Zhang Y, Dipple KM, Vilain E, Huang BL, Finlayson G, Therrell BL, et al. AluY insertion (IVS4-52ins316alu) in the glycerol kinase gene from an individual with benign glycerol kinase deficiency. Hum Mutat [Internet]. 2000;15(4):316–23. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10737976

126. den Hollander a I, ten Brink JB, de Kok YJ, van Soest S, van den Born LI, van Driel M a, et al. Mutations in a human homologue of Drosophila crumbs cause retinitis pigmentosa (RP12). Nat Genet. 1999;23(2):217–21.

127. Beauchamp NJ, Makris M, Preston FE, Peake IR, Daly ME. Major structural defects in the antithrombin gene in four families with type I antithrombin deficiency--partial/complete deletions and rearrangement of the antithrombin gene. Thromb Haemost [Internet]. 2000;83(5):715–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10823268

128. Kloor M, Sutter C, Wentzensen N, Cremer FW, Buckowitz A, Keller M, et al. A large MSH2 Alu insertion mutation causes HNPCC in a German kindred. Hum Genet [Internet]. 2004;115(5):432–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15340835\nhttp://link.springer.com/content/pdf/10.1007/s00439-004-1176-9

129. Muratani K, Hada T, Yamamoto Y, Kaneko T, Shigeto Y, Ohue T, et al. Inactivation of the cholinesterase gene by Alu insertion: possible mechanism for human gene transposition. Proc Natl Acad Sci U S A [Internet]. 1991;88(24):11315–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=53125&tool=pmcentrez&rendertype=abstract

130. Janicic N, Pausova Z, Cole DE, Hendy GN. Insertion of an Alu sequence in the Ca(2+)-sensing receptor gene in familial hypocalciuric hypercalcemia and neonatal severe hyperparathyroidism. Am J Hum Genet [Internet]. 1995;56(4):880–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1801194&tool=pmcentrez&rendertype=abstract

131. Sobrier M-L, Netchine I, Heinrichs C, Thibaud N, Vié-Luton M-P, Van Vliet G, et al. Alu-element insertion in the homeodomain of HESX1 and aplasia of the anterior pituitary. Hum Mutat [Internet]. 2005;25(5):503. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15841484

132. Gallus GN, Cardaioli E, Rufa A, Da Pozzo P, Bianchi S, D'Eramo C, et al. Alu-element insertion in an OPA1 intron sequence associated with autosomal dominant optic atrophy. Mol Vis. 2010;16:178–83.

133. Anagnou NP, Economou-Pachnis A, O'Brien SJ, Modi WS, Nienhuis AW, Tsichlis PN. The human homolog of the Moloney leukemia virus integration 2 locus (MLV12) maps to band p14 of chromosome 5. Genomics [Internet]. 1989 Aug;5(2):354–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2793187

134. Halling KC, Lazzaro CR, Honchel R, Bufill JA, Powell SM, Arndt CA, et al. Hereditary desmoid disease in a family with a germline Alu I repeat mutation of the APC gene. Hum Hered [Internet]. 1999;49(2):97–102. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10077730

135. Su LK, Steinbach G, Sawyer JC, Hindi M, Ward P a, Lynch PM. Genomic rearrangements of the APC tumor-suppressor gene in familial adenomatous polyposis. Hum Genet [Internet]. 2000;106(1):101–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10982189

136. Tucker BA, Scheetz TE, Mullins RF, DeLuca AP, Hoffmann JM, Johnston RM, et al. Exome sequencing and analysis of induced pluripotent stem cells identify the cilia-related gene male germ cell-associated kinase (MAK) as a cause of retinitis pigmentosa. Proc Natl Acad Sci U S A [Internet]. 2011;108(34):E569–76. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/21825139

137. Manco L, Relvas L, Pinto C. Molecular characterization of five Portuguese patients with pyrimidine 5’-nucleotidase deficient hemolytic anemia showing three new P5'NI mutations. … [Internet]. 2006;91(2):2–3. Available from: http://www.haematologica.it/content/91/2/266.short

138. Chen J-M, Masson E, Macek M, Raguénès O, Piskackova T, Fercot B, et al. Detection of two Alu insertions in the CFTR gene. J Cyst Fibros [Internet]. 2008;7(1):37–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17531547

139. Abdelhak S, Kalatzis V, Heilig R, Compain S, Samson D, Vincent C, et al. Clustering of mutations responsible for branchio-oto-renal (BOR) syndrome in the eyes absent homologous region (eyaHR) of EYA1. Hum Mol Genet. 1997;6(13):2247–55.

140. Udaka T, Okamoto N, Aramaki M, Torii C, Kosaki R, Hosokai N, et al. An Alu retrotransposition-mediated deletion of CHD7 in a patient with CHARGE syndrome. Am J Med Genet A [Internet]. 2007;143(7):721–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17334995

141. Bouchet C, Vuillaumier-Barrot S, Gonzales M, Boukari S, Bizec C Le, Fallet C, et al. Detection of an Alu insertion in the POMT1 gene from three French Walker Warburg syndrome families. Mol Genet Metab [Internet]. 2007;90(1):93–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17079174

142. Oldridge M, Zackai EH, McDonald-McGinn DM, Iseki S, Morriss-Kay GM, Twigg SR, et al. De novo alu-element insertions in FGFR2 identify a distinct pathological basis for Apert syndrome. Am J Hum Genet. 1999;64(2):446–61.

143. Bochukova EG, Roscioli T, Hedges DJ, Taylor IB, Johnson D, David DJ, et al. Rare mutations of FGFR2 causing apert syndrome: Identification of the first partial gene deletion, and an Alu element insertion from a new subfamily. Hum Mutat. 2009;30(2):204–11.

144. Tighe PJ, Stevens SE, Dempsey S, Le Deist F, Rieux-Laucat F, Edgar JDM. Inactivation of the Fas gene by Alu insertion: retrotransposition in an intron causing splicing variation and autoimmune lymphoproliferative syndrome. Genes Immun [Internet]. 2002;3 Suppl 1:S66–70. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12215906

145. Stoppa-Lyonnet D, Carter PE, Meo T, Tosi M. Clusters of intragenic Alu repeats predispose the human C1 inhibitor locus to deleterious rearrangements. Proc Natl Acad Sci U S A [Internet]. 1990;87(4):1551–5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=53513&tool=pmcentrez&rendertype=abstract

146. Mustajoki S, Ahola H, Mustajoki P, Kauppinen R. Insertion of Alu element responsible for acute intermittent porphyria. Hum Mutat [Internet]. 1999;13(6):431–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10408772

147. Tappino B, Regis S, Corsolini F, Filocamo M. An Alu insertion in compound heterozygosity with a microduplication in GNPTAB gene underlies Mucolipidosis II. Mol Genet Metab [Internet]. 2008;93(2):129–33. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17964840

148. Miki Y, Katagiri T, Kasumi F, Yoshimoto T, Nakamura Y. Mutation analysis in the BRCA2 gene in primary breast cancers. Nat Genet [Internet]. 1996;13(2):245–7. Available from: http://www.nature.com/ng/journal/v13/n2/abs/ng0696-

245.html\nhttp://www.nature.com/ng/journal/v13/n2/pdf/ng0696-245.pdf

149. Teugels E, De Brakeleer S, Goelen G, Lissens W, Sermijn E, De Grève J. De novo Alu element insertions targeted to a sequence common to the BRCA1 and BRCA2 genes. Hum Mutat. 2005;26(3):284.

150. Schollen E, Keldermans L, Foulquier F, Briones P, Chabas A, Sánchez-Valverde F, et al. Characterization of two unusual truncating PMM2 mutations in two CDG-Ia patients. Mol Genet Metab. 2007;90(4):408–13.

151. Peixoto A, Pinheiro M, Massena L, Santos C, Pinto P, Rocha P, et al. Genomic characterization of two large Alu-mediated rearrangements of the BRCA1 gene. J Hum Genet [Internet]. 2013 Feb;58(2):78–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23223007

152. Wallace MR, Andersen LB, Saulino a M, Gregory PE, Glover TW, Collins FS. A de novo Alu insertion results in neurofibromatosis type 1. Nature [Internet]. 1991;353(6347):864–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1719426

153. Wimmer K, Callens T, Wernstedt A, Messiaen L. The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. PLoS Genet [Internet]. 2011;7(11):e1002371. Available from: http://dx.plos.org/10.1371/journal.pgen.1002371\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3219598&tool=pmcentrez&rendertype=abstract

154. Meischl C, Boer M, Ahlin a, Roos D. A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. Eur J Hum Genet. 2000;8(9):697–703.

155. Brouha B, Meischl C, Ostertag E, de Boer M, Zhang Y, Neijens H, et al. Evidence consistent with human L1 retrotransposition in maternal meiosis I. American journal of human genetics. 2002. p. 327–36.

156. Musova Z, Hedvicakova P, Mohrmann M, Tesarova M, Krepelova A, Zeman J, et al. A novel insertion of a rearranged L1 element in exon 44 of the dystrophin gene: further evidence for possible bias in retroposon integration. Biochem Biophys Res Commun [Internet]. 2006;347(1):145–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16808900

157. Narita N, Nishio H, Kitoh Y, Ishikawa Y, Ishikawa Y, Minami R, et al. Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. J Clin Invest [Internet]. 1993 May;91(5):1862–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8387534

158. Holmes SE, Dombroski BA, Krebs CM, Boehm CD, Kazazian HH. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. Nat Genet. 1994;7(2):143–8.

159. Mukherjee S, Mukhopadhyay a, Banerjee D, Chandak GR, Ray K. Molecular pathology of haemophilia B: identification of five novel mutations including a LINE 1 insertion in Indian patients. Haemophilia [Internet]. 2004;10(3):259–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15086324

160. Morisada N, Rendtorff ND, Nozu K, Morishita T, Miyakawa T, Matsumoto T, et al. Branchio-oto-renal syndrome caused by partial EYA1 deletion due to LINE-1 insertion. Pediatr Nephrol. 2010;25(7):1343–8.

161. Kondo-Iida E, Kobayashi K, Watanabe M, Sasaki J, Kumagai T, Koide H, et al. Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD). Hum Mol Genet [Internet]. 1999;8(12):2303–9. Available from: papers2://publication/uuid/BB3FE7F3-451E-49AD-AA0D-CA4434BB506C

162. Bernard V, Minnerop M, Bürk K, Kreuz F, Gillessen-Kaesbach G, Zühlke C. Exon deletions and intragenic insertions are not rare in ataxia with oculomotor apraxia 2. BMC Med Genet [Internet]. 2009;10:87. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2749023&tool=pmcentrez&rendertype=abstract

163. Lanikova L, Kucerova J, Indrak K, Divoka M, Issa J-P, Papayannopoulou T, et al. β-Thalassemia due to intronic LINE-1 insertion in the β-globin gene (HBB): molecular mechanisms underlying reduced transcript levels of the β-globin(L1) allele. Hum Mutat [Internet]. 2013 Oct;34(10):1361–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23878091

164. Miné M, Chen J-M, Brivet M, Desguerre I, Marchant D, de Lonlay P, et al. A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. Hum Mutat [Internet]. 2007 Feb;28(2):137–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17152059

165. Kagawa T, Oka A, Kobayashi Y, Hiasa Y, Kitamura T, Sakugawa H, et al. Recessive inheritance of population-specific intronic LINE-1 insertion causes a rotor syndrome phenotype. Hum Mutat [Internet]. 2015 Mar;36(3):327–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25546334

166. Nakamura Y, Murata M, Takagi Y, Kozuka T, Nakata Y, Hasebe R, et al. SVA retrotransposition in exon 6 of the coagulation factor IX gene causing severe hemophilia B. Int J Hematol [Internet]. 2015 Jul;102(1):134–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25739383

167. Makino S, Kaji R, Ando S, Tomizawa M, Yasuno K, Goto S, et al. Reduced neuron-specific expression of the TAF1 gene is associated with X-linked dystonia-parkinsonism. Am J Hum Genet. 2007;80(3):393–406.

168. Arca M, Zuliani G, Wilund KR, Campagna F, Fellin R, Bertolini S, et al. Autosomal recessive hypercholesterolaemia in Sardinia, Italy, and mutations in ARH: a clinical and molecular genetic analysis. Lancet [Internet]. 2002;359(9309):841–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11897284

169. Takasu M, Hayashi R, Maruya E, Ota M, Imura K, Kougo K, et al. Deletion of entire HLA-A gene accompanied by an insertion of a retrotransposon. Tissue Antigens [Internet]. 2007;70(2):144–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17610419

170. Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, et al. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. Nature. 1998;394(6691):388–92.

171. Taniguchi-Ikeda M, Kobayashi K, Kanagawa M, Yu C, Mori K, Oda T, et al. Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. Nature. 2011. p. 127–31.

172. Akman HO, Davidzon G, Tanji K, Macdermott EJ, Larsen L, Davidson MM, et al.

Neutral lipid storage disease with subclinical myopathy due to a retrotransposal insertion in the PNPLA2 gene. Neuromuscul Disord [Internet]. 2010;20(6):397–402. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20471263

173. Segal Y, Peissel B, Renieri A, de Marchi M, Ballabio A, Pei Y, et al. LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. Am J Hum Genet. 1999;64(1):62–9.

174. Wang T, Lerer I, Gueta Z, Sagi M, Kadouri L, Peretz T, et al. A deletion/insertion mutation in the BRCA2 gene in a breast cancer family: a possible role of the Alu-polyA tail in the evolution of the deletion. Genes Chromosom Cancer. 2001;31(1):91–5.

175. Qian YDM-DTJHCCCD, Roa JHMRBCKRBBB. Identification of retrotransposons insertion mutations in hereditary cancer. In Myriad Genetics, Inc., Salt Lake City, UT; 2015.

176. Callinan A, Batzer MA, Callinan P a. Retrotransposable Elements and Human Disease. In: Genome Dynamics [Internet]. 2006. p. 104–15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18724056

177. de Boer M, van Leeuwen K, Geissler J, Weemaes CM, van den Berg TK, Kuijpers TW, et al. Primary Immunodeficiency Caused by an Exonized Retroposed Gene Copy Inserted in the CYBB Gene. Hum Mutat. 2014;35(4):486–96.

178. Vogt J, Bengesser K, Claes KB, Wimmer K, Mautner V-F, van Minkelen R, et al. SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. Genome Biol [Internet]. 2014;15(6):R80. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/24958239

179. Van Den Hurk JAJM, Van De Pol DJR, Wissinger B, Van Driel MA, Hoefsloot LH, De Wijs IJ, et al. Novel types of mutation in the choroideremia (CHM) gene: A full-length L1 insertion and an intronic mutation activating a cryptic exon. Hum Genet. 2003;113(3):268–75.

180. Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, et al. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res. 2012;40(4):1666–83.

181. Goodier JL, Ostertag EM, Kazazian HH. Transduction of 3'-flanking sequences is common in L1 retrotransposition. Hum Mol Genet. 2000;9(4):653–7.

182. Pickeral OK, Makałowski W, Boguski MS, Boeke JD. Frequent human genomic DNA transduction driven by line-1 retrotransposition. Genome Res. 2000;10(4):411–5.

183. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. Molecular archeology of L1 insertions in the human genome. Genome Biol. 2002;3(10):research0052.

184. Solyom S, Ewing AD, Hancks DC, Takeshima Y, Awano H, Matsuo M, et al. Pathogenic orphan transduction created by a nonreference LINE-1 retrotransposon. Hum Mutat. 2012;33(2):369–71.

185. Awano H, Malueka RG, Yagi M, Okizuka Y, Takeshima Y, Matsuo M. Contemporary retrotransposition of a novel non-coding gene induces exon-skipping in dystrophin mRNA. Journal of human genetics. 2010. p. 785–90.

186. Moran J V, DeBerardinis RJ, Kazazian HH. Exon shuffling by L1 retrotransposition.

Science. 1999;283(5407):1530–4.

187. Wheelan SJ, Aizawa Y, Han JS, Boeke JD. Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. Genome Res. 2005;15(8):1073–8.

188. Varon R, Gooding R, Steglich C, Marns L, Tang H, Angelicheva D, et al. Partial deficiency of the C-terminal-domain phosphatase of RNA polymerase II is associated with congenital cataracts facial dysmorphism neuropathy syndrome. Nat Genet. 2003;35(2):185–9.

189. Meili D, Kralovicova J, Zagalak J, Bonafe L, Fiori L, Blau N, et al. Disease-causing mutations improving the branch site and polypyrimidine tract: Pseudoexon activation of LINE-2 and antisense alu lacking the poly(T)-Tail. Hum Mutat. 2009;30(5):823–31.

190. Kazazian HH. Mobile elements: drivers of genome evolution. Science. 2004;303(5664):1626–32.

191. Hedges DJ, Deininger PL. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res - Fundam Mol Mech Mutagen. 2007;616(1-2):46–59.

192. Belancio VP, Hedges DJ, Deininger P. Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. Genome Res. 2008;18(3):343–58.

193. Goodier JL, Kazazian HH. Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. Cell. 2008. p. 23–35.

194. Lee J, Ha J, Son SY, Han K. Human genomic deletions generated by SVA-associated events. Comp Funct Genomics. 2012;2012.

195. Hill AS, Foot NJ, Chaplin TL, Young BD. The most frequent constitutional translocation

in humans, the t(11;22)(q23;q11) is due to a highly specific alu-mediated recombination. Hum Mol Genet. 2000;9(10):1525–32.

196. Bailey JA, Liu G, Eichler EE. An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet. 2003;73(4):823–34.

197. Flynn EK, Kamat A, Lach FP, Donovan FX, Kimble DC, Narisu N, et al. Comprehensive Analysis of Pathogenic Deletion Variants in Fanconi Anemia Genes. Hum Mutat [Internet]. 2014;n/a – n/a. Available from: http://doi.wiley.com/10.1002/humu.22680

198. Temtamy S a., Aglan MS, Valencia M, Cocchi G, Pacheco M, Ashour AM, et al. Long interspersed nuclear element-1 (LINE1)-mediated deletion of EVC, EVC2, C4orf6, and STK32B in ellis-van Creveld syndrome with borderline intelligence. Hum Mutat. 2008;29(7):931–8.

199. Burwinkel B, Kilimann MW. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. J Mol Biol. 1998;277(3):513–7.

200. Wu X, Lu Y, Ding Q, You G, Dai J, Xi X, et al. Characterisation of large F9 deletions in seven unrelated patients with severe haemophilia B. 2014;459–65.

201. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, et al. Landscape of somatic retrotransposition in human cancers. Science [Internet]. 2012 Aug 24;337(6097):967–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22745252

202. Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science [Internet]. 2014;345(6196):1251343. Available from:

http://www.sciencemag.org/cgi/doi/10.1126/science.1251343\nhttp://www.ncbi.nlm.nih.gov/pubmed/25082706

203. Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, et al. Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res. 2012;22(12):2328–38.

204. Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Manda SS, et al. Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res. 2015;

205. Shukla R, Upton KR, Muñoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T, et al. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell. 2013;153(1):101–11.

206. Doucet-O'Hare TT, Rodić N, Sharma R, Darbari I, Abril G, Choi J a., et al. LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. Proc Natl Acad Sci [Internet]. 2015;201502474. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1502474112

207. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, et al. Analysis of somatic retrotransposition in human cancers. BMC Proceedings. 2012. p. O23.

208. Rodić N, Steranka JP, Makohon-Moore A, Moyer A, Shen P, Sharma R, et al. Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. Nat Med [Internet]. 2015;(August). Available from: http://www.nature.com/doifinder/10.1038/nm.3919

209. Rodić N, Sharma R, Sharma R, Zampella J, Dai L, Taylor MS, et al. Long interspersed element-1 protein expression is a hallmark of many human cancers. Am J Pathol.

2014;184(5):1280–6.

210. Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M. Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome Res. 2014;24(7):1053–63.

211. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011. p. 646–74.

212. Hanahan D, Weinberg RA. The Hallmarks of Cancer Review evolve progressively from normalcy via a series of pre. Cell [Internet]. 2000;100(1):57–70. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092-8674(00)81683-9\nhttp://www.sciencedirect.com/science/article/B6WSN-4195FC1-5/2/aef1d48431eadea4567b697b1fee0514

213. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A and HK. Environmental and heritable factors in the causation of cancer-analyses of Cohorts of Twins from Sweden , Denmark , and Finland. N Engl J Med. 2000;343(2):78–85.

214. Sorenson TIA, Nielsen GG, Andersen PK TT. Genetic and Environmental Influences on Premature Death in Adult Adoptees. N Engl J Med. 1988;318(12):727–32.

215. Garcia-Perez JL, Morell M, Scheys JO, Kulpa DA, Morell S, Carter CC, et al. Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature. 2010;466(7307):769–73.

216. Ostertag EM, Prak ET, DeBerardinis RJ, Moran J V, Kazazian HH. Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res. 2000;28(6):1418–23.

217. Garcia-Perez JL, Marchetto MCN, Muotri AR, Coufal NG, Gage FH, O'Shea KS, et al. LINE-1 retrotransposition in human embryonic stem cells. Hum Mol Genet. 2007;16(13):1569–77.

218. Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, et al. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev. 2009;23(11):1303–12.

219. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, et al. L1 retrotransposition in human neural progenitor cells. Nature. 2009;460(7259):1127–31.

220. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011. p. 534–7.

221. Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, et al. Single-neuron sequencing analysis of l1 retrotransposition and somatic mutation in the human brain. Cell. 2012;151(3):483–96.

222. Boffetta P, Nyberg F. Contribution of environmental factors to cancer risk. British Medical Bulletin. 2003. p. 71–94.

223. Fornace AJ, Mitchell JB. Induction of B2 RNA polymerase III transcription by heat shock: Enrichment for heat shock induced sequences in rodent cells by hybridization subtraction. Nucleic Acids Res. 1986;14(14):5793–811.

224. Denissenko MF, Pao A, Tang M PG. Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in p53. Science (80-). 1996;274(5286):430–2.

225. Tabatabaei SM, Heyworth JS, Knuiman MW, Fritschi L. Dietary benzo[a]pyrene intake

from meat and the risk of colorectal cancer. Cancer Epidemiol Biomarkers Prev. 2010;19(12):3182–4.

226. Rathore K, Wang HCR. Mesenchymal and stem-like cell properties targeted in suppression of chronically-induced breast cell carcinogenesis. Cancer Lett [Internet]. Elsevier Ireland Ltd; 2013;333(1):113–23. Available from: http://dx.doi.org/10.1016/j.canlet.2013.01.030

227. Stribinskis V, Ramos KS. Activation of human long interspersed nuclear element 1 retrotransposition by benzo(a)pyrene, an ubiquitous environmental carcinogen. Cancer Res. 2006;66(5):2616–20.

228. Beveridge R, Pintos J, Parent M, Asselin J, Siemiatycki J. Lung Cancer Risk Associated With Occupational Exposure to Nickel , Chromium VI , and Cadmium in Two Population-Based Case – Control Studies in Montreal. Am J Ind Med. 2010;485(53):476–85.

229. Al-Qubaisi MS, Rasedee A, Flaifel MH, Ahmad SHJ, Hussein-Al-Ali S, Hussein MZ, et al. Cytotoxicity of nickel zinc ferrite nanoparticles on cancer cells of epithelial origin. Int J Nanomedicine. 2013;8:2497–508.

230. Kale SP, Moore L, Deininger PL, Roy-Engel AM. Heavy metals stimulate human LINE-1 retrotransposition. In: International Journal of Environmental Research and Public Health. 2005. p. 14–23.

231. Toyokuni S, Okamoto K, Yodoi J, Hiai H. Persistent oxidative stress in cancer. FEBS Lett. 1995;358(1):1–3.

232. Giorgi G, Marcantonio P, Del Re B. LINE-1 retrotransposition in human neuroblastoma

cells is affected by oxidative stress. Cell Tissue Res. 2011;346(3):383–91.

233. Maxwell PH, Burhans WC, Curcio MJ. Retrotransposition is associated with genome instability during chronological aging. Proc Natl Acad Sci U S A [Internet]. 2011;108(51):20376–81. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3251071&tool=pmcentrez&rendertype=abstract

234. Jacobs KB, Yeager M, Zhou W, Wacholder S, Wang Z, Rodriguez-Santiago B, et al. Detectable clonal mosaicism and its relationship to aging and cancer. Nat Genet. 2012;44(6):651–8.

235. Laurie CC, Laurie C a, Rice K, Doheny KF, Zelnick LR, McHugh CP, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. Nat Genet [Internet]. Nature Publishing Group; 2012;44(6):642–50. Available from: http://dx.doi.org/10.1038/ng.2271

236. Carreira PE, Richardson SR, Faulkner GJ. L1 retrotransposons, cancer stem cells and oncogenesis. FEBS Journal. 2014. p. 63–73.

237. Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. Cancer Res. 1992;52(3):643–5.

238. Kinzler KW, Nilbert MC, Su LK, Vogelstein B, Bryan TM, Levy DB, et al. Identification of FAP locus genes from chromosome 5q21. Science. 1991;253(5020):661–5.

239. Kinzler KW, Nilbert MC, Vogelstein B, Bryan TM, Levy DB, Smith KJ, et al. Identification of a gene located at chromosome 5q21 that is mutated in colorectal cancers.

Science (80-) [Internet]. 1991;251(4999):1366–70. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1848370

240. Ray DA, Batzer MA. Reading TE leaves: New approaches to the identification of transposable element insertions. Genome Research. 2011. p. 813–20.

241. Faulkner GJ. Retrotransposons: Mobile and mutagenic from conception to death. FEBS Lett [Internet]. Federation of European Biochemical Societies; 2011;585(11):1589–94. Available from: http://dx.doi.org/10.1016/j.febslet.2011.03.061

242. Ovchinnikov I, Troxel AB, Swergold GD. Genomic characterization of recent human LINE-1 insertions: Evidence supporting random insertion. Genome Res. 2001;11(12):2050–8.

243. Badge RM, Alisch RS, Moran J V. ATLAS: a system to selectively identify human-specific L1 insertions. Am J Hum Genet. 2003;72(4):823–38.

244. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, et al. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell. 2010;141(7):1253–61.

245. Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran J V, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005;435(7044):903–10.

246. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, et al. L1 retrotransposition in human neural progenitor cells. Nature [Internet]. 2009;460(7259):1127–31. Available from: http://dx.doi.org/10.1038/nature08248

247. Hastings PJ, Ira G, Lupski JR. A microhomology-mediated break-induced replication model for the origin of human copy number variation. PLoS Genetics. 2009.

248. Graham T, Boissinot S. The genomic distribution of L1 elements: The role of insertion bias and natural selection. Journal of Biomedicine and Biotechnology. 2006.

249. Ewing AD, Kazazian HH. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. 2010;20(9):1262–70.

250. Cai X, Evrony GD, Lehmann HS, Elhosary PC, Mehta BK, Poduri A, et al. Single-Cell, Genome-wide Sequencing Identifies Clonal Somatic Copy-Number Variation in the Human Brain. Cell Rep [Internet]. Elsevier; 2014;8(5):1280–9. Available from: http://dx.doi.org/10.1016/j.celrep.2014.07.043

251. Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, et al. Mobile elements create structural variation: Analysis of a complete human genome. Genome Res. 2009;19(9):1516–26.

252. Hormozdiari F, Konkel MK, Prado-Martinez J, Chiatante G, Herraez IH, Walker J a, et al. Rates and patterns of great ape retrotransposition. Proc Natl Acad Sci U S A [Internet]. 2013;110(33):13457–62. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3746892&tool=pmcentrez&rendertype=abstract

253. Huang CRL, Schneider AM, Lu Y, Niranjan T, Shen P, Robinson M a., et al. Mobile interspersed repeats are major structural variants in the human genome. Cell [Internet]. Elsevier Ltd; 2010;141(7):1171–82. Available from:

http://dx.doi.org/10.1016/j.cell.2010.05.026

254. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, et al. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet. 2002;31(2):159–65.

255. Paterson AL, Weaver JMJ, Eldridge MD, Tavaré S, Fitzgerald RC, Edwards P a. W. Mobile element insertions are frequent in oesophageal adenocarcinomas and can mislead paired-end sequencing analysis. BMC Genomics [Internet]. BMC Genomics; 2015;16(1):473. Available from: http://www.biomedcentral.com/1471-2164/16/473

256. Rodić N, Sharma R, Sharma R, Zampella J, Dai L, Taylor MS, et al. Long interspersed element-1 protein expression is a hallmark of many human cancers. Am J Pathol. 2014;184(5):1280–6.

257. Badillo R, Francis D. Diagnosis and treatment of gastroesophageal reflux disease. World J Gastrointest Pharmacol Ther [Internet]. 2014;5(3):105–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25133039

258. Barrett N. Chronic peptic ulcer of the oesophagus and "oesophagitis." Br J Surg. 1950;38(150):175–82.

259. Sappati Biyyani RS, Chak A. Barrett's esophagus: review of diagnosis and treatment. Gastroenterol Rep [Internet]. 2013;1(1):9–18. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3941437&tool=pmcentrez&rendertype=abstract

260. Modiano N, Gerson LB. Barrett's esophagus: Incidence, etiology, pathophysiology, prevention and treatment. Therapeutics and Clinical Risk Management. 2007. p. 1035–45.

261. Raskind WH, Norwood T, Levine DS, Haggitt RC, Rabinovitch PS, Reid BJ. Persistent clonal areas and clonal expansion in Barrett's esophagus. Cancer Res. 1992;52(10):2946–50.

262. Maley CC, Galipeau PC, Li X, Sanchez CA, Paulson TG, Reid BJ. Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. Cancer Res. 2004;64(10):3414–27.

263. Van der Klift HM, Tops CM, Hes FJ, Devilee P, Wijnen JT. Insertion of an SVA element, a nonautonomous retrotransposon, in PMS2 intron 7 as a novel cause of lynch syndrome. Hum Mutat. 2012;33(7):1051–5.

264. Goodier JL. Retrotransposition in tumors and brains. Mob DNA [Internet]. 2014;5(1):11. Available from: http://www.mobilednajournal.com/content/5/1/11

265. Gilbert N, Lutz-Prigge S, Moran J V. Genomic deletions created upon LINE-1 retrotransposition. Cell. 2002;110(3):315–25.

266. Bratthauer GL, Cardiff RD, Fanning TG. Expression of LINE-1 retrotransposons in human breast cancer. Cancer. 1994;73(9):2333–6.

267. Bratthauer GL, Fanning TG. LINE-1 retrotransposon expression in pediatric germ cell tumors. Cancer. 1993;71(7):2383–6.

268. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, et al. Ubiquitous L1 Mosaicism in Hippocampal Neurons. Cell [Internet]. 2015;161(2):228–39. Available from: http://linkinghub.elsevier.com/retrieve/pii/S009286741500313X

269. Abedi-Ardekani B, Hainaut P. Cancers of the upper gastro-intestinal tract: A review of somatic mutation distributions. Arch Iran Med. 2014;17(4):286–92.

270. Ohashi S, Miyamoto S, Kikuchi O, Goto T, Amanuma Y, Muto M. Recent Advances from Basic and Clinical Studies of Esophageal Squamous Cell Carcinoma. Gastroenterology [Internet]. Elsevier Ltd; 2015; Available from: http://linkinghub.elsevier.com/retrieve/pii/S0016508515013104

271. Kamangar F, Malekzadeh R, Dawsey SM, Saidi F. Esophageal cancer in Northeastern Iran: a review. Arch Iran Med [Internet]. 2007;10(1):70–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17198458

272. Enzinger PC, Mayer RJ. Esophageal cancer. N Engl J Med [Internet]. 2003;349:2241–52. Available from: <Go to ISI>://WOS:A1994PX97100001

273. Lambert R, Hainaut P. Epidemiology of oesophagogastric cancer. Best Pract Res Clin Gastroenterol [Internet]. 2007;21(6):921–45. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1521691807001060

274. Islami F, Kamangar F, Nasrollahzadeh D, Møller H, Boffetta P, Malekzadeh R. Oesophageal cancer in Golestan Province, a high-incidence area in northern Iran - A review. Eur J Cancer [Internet]. 2009;45(18):3156–65. Available from: http://dx.doi.org/10.1016/j.ejca.2009.09.018

275. Tran GD, Sun X-D, Abnet CC, Fan J-H, Dawsey SM, Dong Z-W, et al. Prospective study of risk factors for esophageal and gastric cancers in the Linxian general population trial cohort in China. Int J Cancer [Internet]. 2005;113(3):456–63. Available from: http://doi.wiley.com/10.1002/ijc.20616

276. Baba Y, Watanabe M, Murata A, Shigaki H, Miyake K, Ishimoto T, et al. LINE-1 Hypomethylation, DNA Copy Number Alterations, and CDK6 Amplification in

Esophageal Squamous Cell Carcinoma. Clin Cancer Res [Internet]. 2014;20(5):1114–24. Available from: http://clincancerres.aacrjournals.org/cgi/doi/10.1158/1078-0432.CCR-13-1645

277. Goodier JL, Pereira GC, Cheung LE, Rose RJ, Kazazian HH. The Broad-Spectrum Antiviral Protein ZAP Restricts Human Retrotransposition. PLOS Genet [Internet]. 2015;11(5):e1005252. Available from: http://dx.plos.org/10.1371/journal.pgen.1005252

278. Zhao K, Du J, Han X, Goodier JL, Li P, Zhou X, et al. Modulation of LINE-1 and Alu/SVA Retrotransposition by Aicardi-Goutières Syndrome-Related SAMHD1. Cell Rep. 2013;4(6):1108–15.

279. Baba Y, Murata A, Watanabe M, Baba H. Clinical implications of the LINE-1 methylation levels in patients with gastrointestinal cancer. Surg Today. 2013;(23689061):1–10.

280. Irahara N, Nosho K, Baba Y, Shima K, Lindeman NI, Hazra A, et al. Precision of pyrosequencing assay to measure LINE-1 methylation in colon cancer, normal colonic mucosa, and peripheral blood cells. J Mol Diagn. 2010;12(2):177–83.

281. Iwagami S, Baba Y, Watanabe M, Shigaki H, Miyake K, Ishimoto T, et al. LINE-1 Hypomethylation Is Associated With a Poor Prognosis Among Patients With Curatively Resected Esophageal Squamous Cell Carcinoma. Ann Surg [Internet]. 2013;257(3):449–55. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00000658-201303000-00012

282. Martin SL. The ORF1 protein encoded by LINE-1: Structure and function during L1

retrotransposition. Journal of Biomedicine and Biotechnology. 2006.

283. Zhang H-Z, Jin G-F, Shen H-B. Epidemiologic differences in esophageal cancer between Asian and Western populations. Chin J Cancer [Internet]. 2012;31(6):281–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3777490&tool=pmcentrez&rendertype=abstract

284. Morrish TA, Garcia-Perez JL, Stamato TD, Taccioli GE, Sekiguchi J, Moran J V. Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature. 2007;446(7132):208–12.

285. An O, Dall'Olio GM, Mourikis TP, Ciccarelli FD. NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. Nucleic Acids Res [Internet]. 2015 Oct 29;gkv1123. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26516186

286. Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, et al. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature. 2010;466(7308):869–73.

287. Dulak AM, Stojanov P, Peng S, Lawrence MS, Fox C, Stewart C, et al. Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. Nat Genet [Internet]. Nature Publishing Group; 2013;45(5):478–86. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3678719&tool=pmcentrez&rendertype=abstract

288. Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011;333(6046):1157–60.

289. Jones S, Zhang X, Parsons DW, Lin JC-H, Leary RJ, Angenendt P, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science [Internet]. 2008;321(5897):1801–6. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2848990&tool=pmcentrez&rendertype=abstract

290. Cao Y, He M, Gao Z, Peng Y, Li Y, Li L, et al. Activating hotspot L205R mutation in PRKACA and adrenal Cushing's syndrome. Science [Internet]. 2014;344(6186):913–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24700472

291. Fleming JL, Dworkin AM, Allain DC, Fernandez S, Wei L, Peters SB, et al. Allele-specific imbalance mapping identifies HDAC9 as a candidate gene for cutaneous squamous cell carcinoma. Int J Cancer. 2014;134(1):244–8.

292. Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature [Internet]. 2008;455(7216):1069–75. Available from: http://www.nature.com/doifinder/10.1038/nature07423

293. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61–70.

294. Wang L, Tsutsumi S, Kawaguchi T, Nagasaki K, Tatsuno K, Yamamoto S, et al. Whole-

exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. Genome Res. 2012;22(2):208–19.

295. Bonne A, Vreede L, Kuiper RP, Bodmer D, Jansen C, Eleveld M, et al. Mapping of constitutional translocation breakpoints in renal cell cancer patients: identification ofKCNIP4as a candidate gene. Cancer Genet Cytogenet [Internet]. 2007;179(1):11–8. Available from: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=17981209&retmode=ref&cmd=prlinks

296. Wang D, Wang L, Zhou J, Pan J, Qian W, Fu J, et al. Reduced Expression of PTPRD Correlates with Poor Prognosis in Gastric Adenocarcinoma. PLoS One [Internet]. 2014;9(11):e113754. Available from: http://dx.plos.org/10.1371/journal.pone.0113754

297. Lauc G, Huffman JE, Pučić M, Zgaga L, Adamczyk B, Mužinić A, et al. Loci Associated with N-Glycosylation of Human Immunoglobulin G Show Pleiotropy with Autoimmune Diseases and Haematological Cancers. PLoS Genet [Internet]. 2013;9(1):e1003225. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3561084&tool=pmcentrez&rendertype=abstract

298. Cheung K-F, Lam CNY, Wu K, Ng EKO, Chong WWS, Cheng ASL, et al. Characterization of the gene structure, functional significance, and clinical application of RNF180, a novel gene in gastric cancer. Cancer [Internet]. 2012;118:947–59. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21717426

299. Deng J, Liang H, Ying G, Zhang R, Wang B, Yu J, et al. Methylation of CpG sites in

RNF180 DNA promoter prediction poor survival of gastric cancer. Oncotarget [Internet]. 2014;5(10):3173–83. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4102801&tool=pmcentrez&rendertype=abstract

300. Vazquez A, Kulkarni D, Grochola LF, Bond GL, Barnard N, Toppmeyer D, et al. A genetic variant in a PP2A regulatory subunit encoded by the PPP2R2B gene associates with altered breast cancer risk and recurrence. Int J Cancer [Internet]. 2011 May 15;128(10):2335–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20669227

301. Tan J, Lee PL, Li Z, Jiang X, Lim YC, Hooi SC, et al. B55β-associated PP2A complex controls PDK1-directed myc signaling and modulates rapamycin sensitivity in colorectal cancer. Cancer Cell [Internet]. Elsevier Inc.; 2010 Nov 16;18(5):459–71. Available from: http://dx.doi.org/10.1016/j.ccr.2010.10.021

302. Wen D, Xu Z, Xia L, Liu X, Tu Y, Lei H, et al. Important Role of SUMOylation of Spliceosome Factors in Prostate Cancer Cells. J Proteome Res [Internet]. 2014;13(8):3571–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25027693

303. Bignone P a, Lee KY, Liu Y, Emilion G, Finch J, Soosay a ER, et al. RPS6KA2, a putative tumour suppressor gene at 6q27 in sporadic epithelial ovarian cancer. Oncogene. 2007;26(May 2006):683–700.

304. Slattery ML, Lundgreen A, Herrick JS, Wolff RK. Genetic variation in RPS6KA1, RPS6KA2, RPS6KB1, RPS6KB2, and PDK1 and risk of colon or rectal cancer. Mutat Res - Fundam Mol Mech Mutagen. 2011;706(1-2):13–20.

305. Serra V, Eichhorn PJ a, García-García C, Ibrahim YH, Prudkin L, Sánchez G, et al.

RSK3/4 mediate resistance to PI3K pathway inhibitors in breast cancer. J Clin Invest [Internet]. 2013;123(6):2551–63. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3668839&tool=pmcentrez&rendertype=abstract

306. Milosevic N, Kühnemuth B, Mühlberg L, Ripka S, Griesmann H, Lölkes C, et al. Synthetic Lethality Screen Identifies RPS6KA2 as Modifier of Epidermal Growth Factor Receptor Activity in Pancreatic Cancer. Neoplasia [Internet]. 2013;15(12):1354–62. Available from: http://www.sciencedirect.com/science/article/pii/S1476558613800048

307. Fernandez-Rozadilla C, Cazier J-B, Tomlinson IP, Carvajal-Carmona LG, Palles C, Lamas MJ, et al. A colorectal cancer genome-wide association study in a Spanish cohort identifies two variants associated with colorectal cancer risk at 1p33 and 8p12. BMC Genomics [Internet]. 2013;14:55. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3616862&tool=pmcentrez&rendertype=abstract

308. Kerns SL, Stone NN, Stock RG, Rath L, Ostrer H, Rosenstein BS. A 2-stage genome-wide association study to identify single nucleotide polymorphisms associated with development of urinary symptoms after radiotherapy for prostate cancer. J Urol [Internet]. 2013 Jul;190(1):102–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23376709

309. Mukhopadhyay D, Jung J, Murmu N, Houchen CW, Dieckgraefe BK, Anant S. CUGBP2 plays a critical role in apoptosis of breast cancer cells in response to genotoxic injury. Ann N Y Acad Sci [Internet]. 2003 Dec;1010:504–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15033780

310. Natarajan G, Ramalingam S, Ramachandran I, May R, Queimado L, Houchen CW, et al.

CUGBP2 downregulation by prostaglandin E2 protects colon cancer cells from radiation-induced mitotic catastrophe. Am J Physiol Gastrointest Liver Physiol [Internet]. 2008;294(5):G1235–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18325984

311. Matta A, Tripathi SC, DeSouza L V., Grigull J, Kaur J, Chauhan SS, et al. Heterogeneous ribonucleoprotein K is a marker of oral leukoplakia and correlates with poor prognosis of squamous cell carcinoma. Int J Cancer. 2009;125(6):1398–406.

312. Kitamura K, Seike M, Okano T, Matsuda K, Miyanaga A, Mizutani H, et al. MiR-134/487b/655 cluster regulates TGF-β-induced epithelial-mesenchymal transition and drug resistance to gefitinib by targeting MAGI2 in lung adenocarcinoma cells. Mol Cancer Ther [Internet]. 2014;13(2):444–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24258346

313. Sun P-H, Ye L, Mason MD, Jiang WG. Protein tyrosine phosphatase μ (PTP μ or PTPRM), a negative regulator of proliferation and invasion of breast cancer cells, is associated with disease prognosis. PLoS One [Internet]. 2012;7(11):e50183. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3502354&tool=pmcentrez&rendertype=abstract

314. Burgoyne AM, Phillips-Mason PJ, Burden-Gulley SM, Robinson S, Sloan AE, Miller RH, et al. Proteolytic cleavage of protein tyrosine phosphatase mu regulates glioblastoma cell migration. Cancer Res [Internet]. 2009;69(17):6960–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19690139

315. Timson G, Banavali S, Gutierrez MI, Magrath I, Bhatia KG, Goyns MH. High level expression of N-acetylgluosamine-6-O-sulfotransferase is characteristic of a subgroup of

paediatric precursor-B acute lymphoblastic leukaemia. Cancer Lett [Internet]. Elsevier Ireland Ltd; 2006;242(2):239–44. Available from: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=16386360

316. Vargas PA, Speight PM, Bingle C, Barrett AW, Bingle L. Expression of Plunc Family Members in Benign and Malignant Salivary Gland Tumours. Oral Dis [Internet]. 2008;14(7):613–9. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2853704/\nhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC2853704/pdf/ukmss-29166.pdf

317. González-Arriagada WA, Ramos LMA, Silva AA, Vargas PA, Coletta R Della, Bingle L, et al. Salivary BPIFA1 (SPLUNC1) and BPIFA2 (SPLUNC2 A) are modified by head and neck cancer radiotherapy. Oral Surg Oral Med Oral Pathol Oral Radiol [Internet]. Elsevier Inc.; 2015;119(1):48–58. Available from: http://linkinghub.elsevier.com/retrieve/pii/S2212440314012978

318. Köhler S, Ullrich S, Richter U, Schumacher U. E-/P-selectins and colon carcinoma metastasis: first in vivo evidence for their crucial role in a clinically relevant model of spontaneous metastasis formation in the lung. Br J Cancer. 2010;102(3):602–9.

319. Dymicka-Piekarska V, Kemona H. Does colorectal cancer clinical advancement affect adhesion molecules (sP-selectin, sE-selectin and ICAM-1) concentration? Thromb Res [Internet]. Elsevier Ltd; 2009;124(1):80–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19136145

320. Wei M, Tai G, Gao Y, Li N, Huang B, Zhou Y, et al. Modified heparin inhibits P-selectin-mediated cell adhesion of human colon carcinoma cells to immobilized platelets under

dynamic flow conditions. J Biol Chem [Internet]. 2004;279(28):29202–10. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15133030

321. Figueiredo JC, Hsu L, Hutter CM, Lin Y, Campbell PT, Baron JA, et al. Genome-wide diet-gene interaction analyses for risk of colorectal cancer. PLoS Genet [Internet]. 2014;10(4):e1004228. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3990510&tool=pmcentrez&rendertype=abstract

322. Cha J-D, Kim HJ, Cha I-H. Genetic alterations in oral squamous cell carcinoma progression detected by combining array-based comparative genomic hybridization and multiplex ligation-dependent probe amplification. Oral Surgery, Oral Med Oral Pathol Oral Radiol Endodontology [Internet]. Elsevier Inc.; 2011;111(5):594–607. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1079210410009200

323. Davidson B, Abeler VM, Førsund M, Holth A, Yang Y, Kobayashi Y, et al. Gene expression signatures of primary and metastatic uterine leiomyosarcoma. Hum Pathol [Internet]. 2014;45(4):691–700. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965648&tool=pmcentrez&rendertype=abstract

324. Won HH, Lee J, Park JO, Park YS, Lim HY, Kang WK, et al. Polymorphic markers associated with severe oxaliplatin-induced, chronic peripheral neuropathy in colon cancer patients. Cancer. 2012;118(11):2828–36.

325. Li H, Yu B, Li J, Su L, Yan M, Zhu Z, et al. Overexpression of lncRNA H19 enhances carcinogenesis and metastasis of gastric cancer. Oncotarget [Internet]. 2014;5(8):2318–29. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4039165&tool=pmcentrez&rendertype=abstract

326. Vater I, Montesinos-Rongen M, Schlesner M, Haake a, Purschke F, Sprute R, et al. The mutational pattern of primary lymphoma of the central nervous system determined by whole-exome sequencing. Leukemia [Internet]. Nature Publishing Group; 2014;29(3):677–85. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25189415

327. Boeva V, Jouannet S, Daveau R, Combaret V, Pierre-Eugène C, Cazes A, et al. Breakpoint Features of Genomic Rearrangements in Neuroblastoma with Unbalanced Translocations and Chromothripsis. PLoS One [Internet]. 2013;8(8):e72182. Available from: http://dx.plos.org/10.1371/journal.pone.0072182

328. Adélaïde J, Chaffanet M, Mozziconacci MJ, Popovici C, Conte N, Fernandez F, et al. Translocation and coamplification of loci from chromosome arms 8p and 11q in the {MDA-MB-175} mammary carcinoma cell line. Int J Oncol [Internet]. 2000;16(4):683–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10717235

329. Eckel-Passow JE, Serie DJ, Bot BM, Joseph RW, Cheville JC, Parker AS. ANKS1B is a smoking-related molecular alteration in clear cell renal cell carcinoma. BMC Urol [Internet]. BMC Urology; 2014;14(1):14. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3944917&tool=pmcentrez&rendertype=abstract

330. Rose JE, Behm FM, Drgon T, Johnson C, Uhl GR. Personalized smoking cessation: interactions between nicotine dose, dependence and quit-success genotype score. Mol Med [Internet]. 16(7-8):247–53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20379614

331. Uhl GR, Liu Q-R, Drgon T, Johnson C, Walther D, Rose JE, et al. Molecular genetics of successful smoking cessation: convergent genome-wide association study results. Arch Gen Psychiatry [Internet]. 2008;65(6):683–93. Available from: http://archpsyc.jamanetwork.com/article.aspx?articleid=482738

332. Uhl GR, Liu Q-R, Drgon T, Johnson C, Walther D, Rose JE. Molecular genetics of nicotine dependence and abstinence: whole genome association using 520,000 SNPs. BMC Genet. 2007;8:10.

333. Tang H, Wei P, Duell EJ, Risch HA, Olson SH, Bueno-de-Mesquita HB, et al. Axonal guidance signaling pathway interacting with smoking in modifying the risk of pancreatic cancer: a gene- and pathway-based interaction analysis of GWAS data. Carcinogenesis [Internet]. 2014;35(5):1039–45. Available from: http://www.carcin.oxfordjournals.org/cgi/doi/10.1093/carcin/bgu010

334. Uhl GR, Drgon T, Johnson C, Walther D, David SP, Aveyard P, et al. Genome-wide association for smoking cessation success: participants in the Patch in Practice trial of nicotine replacement. Pharmacogenomics [Internet]. 2010 Mar;11(3):357–67. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20235792

335. Dong J, Hu Z, Wu C, Guo H, Zhou B, Lv J, et al. Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. Nat Genet [Internet]. Nature Publishing Group; 2012;44(8):895–9. Available from: http://dx.doi.org/10.1038/ng.2351

336. Tsaprouni LG, Yang T, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Epigenetics Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics [Internet]. 2014;9(10):1382–96. Available

from: http://www.ncbi.nlm.nih.gov/pubmed/25424692

337. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, Luquette LJ, et al. Analysis of somatic retrotransposition in human cancers. BMC Proceedings. 2012. p. O23.

338. Voytas DF, Cummings MP, Koniczny A, Ausubel FM, Rodermel SR. copia-like retrotransposons are ubiquitous among plants. Proc Natl Acad Sci U S A [Internet]. 1992 Aug 1;89(15):7124–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/1379734

339. Mir A a, Philippe C, Cristofari G. euL1db: the European database of L1HS retrotransposon insertions in humans. Nucleic Acids Res [Internet]. 2014;17–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25352549

Tara Theresa Doucet

Curriculum Vitae

Johns Hopkins University
Department of Human Genetics
733 North Broadway, BRB 515
Baltimore, MD 21205

337-298-8558
226 Inlet Drive
Pasadena, MD 21122
tdo87@outlook.com

Education

2010- 2015 **Johns Hopkins University, pursuing a Doctorate of Philosophy in Human Genetics**
Institute of Genetic Medicine, Department of Human Genetics, Baltimore, MD

2006- 2010 **Clemson University, Bachelor of Art, Biology and Minor in Spanish**
Clemson University, Department of Biological Sciences Clemson, SC
GPA: 3.53 Graduated with both Departmental and General Honors.

Graduate Research Experience

2010- 2015 **Thesis Research on Retrotransposons and Cancer**
My projects in the Kazazian lab have revolved around the concept of somatic mosaicism caused by retrotransposons in the human genome, within and between individuals as well as its role in carcinogenesis. I have utilized a variety of techniques to establish the copy number of certain retrotransposon insertions as well as employing next generation sequencing, coupled with a capture targeted to retrotransposons, to assess the activity of these elements in various individuals, diseases, and populations.

Publications

Doucet-O'Hare T, Rodić N, Sharma R, Darbari I, Abril G, Choi JA, Ahn JY, Cheng Y, Anders RA, Burns, KH, Meltzer SJ, Kazazian HH Jr. (2015) LINE-1 Expression and Retrotransposition in Barrett's Esophagus and Esophageal Carcinoma. *Proceedings of the National Academy of Sciences.* September 112(35):E4894-4900. PMID: 26283398

Solyom S, Ewing AD, Rahrmann EP, **Doucet T**, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B, Wheelan S, Upton KR, Shukla R, Faulkner GJ, Largaespada DA, Kazazian HH Jr. (2012) Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Research.* December (12):2328-2338. PMID: 22968929.

Ingram-Smith C, Wharton J, Reinholz C, **Doucet T**, Hesler R, Smith, K. (2015) "The Role of Active Site Residues in ATP Binding and Catalysis in the *Methanosarcina thermophila* Acetate Kinase." *Life* 5, no. 1: 861-871. PMID: 25775277.

Doucet T and Haig H. Kazazian Jr. (2015) "Long interspersed element sequencing (L1-seq): a method to identify somatic LINE-1 insertions in the human genome. *Methods in Molecular Biology.* (In Press)

Doucet-O'Hare T, Sharma R, Rodić N, Burns KH, Anders RA, Kazazian HH Jr. "Frequent sub-clonal LINE-1 insertions in normal esophagus expand in subsequent esophageal squamous cell carcinoma." (submitted to Oncogene December 2015).

Presentations

October 2015	**American Society of Human Genetics Conference (Baltimore)** "LINE-1 activity in and expression in Normal esophagus, Barrett's esophagus, esophageal carcinoma, and Esophageal Squamous cell carcinoma" (poster)
July 2015	**Gordon Research Seminar and Conference on Human Genetics and Genomics (Salve Regina University)** "LINE-1 Activity and Expression in Normal Esophagus, Barrett's Esophagus, and Esophageal Carcinoma" (poster)
June 2015	**FASEB Mobile Elements Conference (West Palm Beach)** "LINE-1 Activity and Expression in Normal Esophagus, Barrett's Esophagus, and Esophageal Carcinoma" (poster)
January 2015	**Partnering Towards Discovery Seminar Series (Johns Hopkins)** "Cancer Biology and the Junk Genome" (oral presentation)
October 2014	**American Society of Human Genetics Conference (San Diego)** "LINE-1 Retrotransposition in Barrett's Esophagus and Esophageal Adenocarcinoma" (poster)
June 2014	**Keystone Conference on Mobile Elements (Santa Fe)** "LINE-1 Retrotransposition in Barrett's Esophagus and Esophageal Carcinoma" (poster)
October 2013	**American Society of Human Genetics Conference (Boston)** "Individual variation in the rate of retrotransposition in iPS cells and its effect on genomic instability with regard to medical utility." (poster)
June 2013	**FASEB Mobile Elements Conference (Big Sky)** "Somatic Variation and Genomic Instability Generated by Retrotransposons in Barrett's Esophagus." (poster)

Academic Honors, Awards, and Leadership Positions

July 2015	Runner-up for best poster at Gordon Seminar on Human Genetics and Genomics
2011-2013	Team Leader in Incentive Mentoring Program
2010- Present	Member of the American Society of Human Genetics

References

Haig H. Kazazian Jr., M.D.
Professor of Pediatrics, Molecular Biology, and Genetics
McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine
733 North Broadway, Baltimore, MD 21205
hkazazi1@jhmi.edu

Kathleen H. Burns, M.D., Ph.D.
Associate Professor of Pathology and Oncology
Johns Hopkins University School of Medicine
720 Rutland Avenue, Ross 524
Baltimore MD 21205
kburns@jhmi.edu

Andy McCallion, PhD.
Associate Professor of Molecular and Comparative Pathobiology, Pediatrics, and Medicine
Assistant Director of Human Genetics Graduate Program
McCusick-Nathans Institute of Genetic Medicine, Johns Hopkins School of Medicine

733 North Broadway, Baltimore, MD 21205
amccall2@jhmi.edu